2

Automatic identification of causal relations in text and their use for improving precision in information retrieval

This is a reformatted version of the original dissertation, and converted from WordPerfect to MS Word to PDF format. The content is the same, but there are some conversion errors.

AUTOMATIC IDENTIFICATION OF CAUSAL RELATIONS IN TEXT AND THEIR USE FOR IMPROVING PRECISION IN INFORMATION RETRIEVAL

Christopher Soo-Guan Khoo

ABSTRACT

This study represents one attempt to make use of relations expressed in text to improve information retrieval effectiveness. In particular, the study investigated whether the information obtained by matching causal relations expressed in documents with the causal relations expressed in users' queries could be used to improve document retrieval results in comparison to using just term matching without considering relations.

An automatic method for identifying and extracting cause-effect information in Wall Street Journal text was developed. The method uses linguistic clues to identify causal relations without recourse to knowledge-based inferencing. The method was successful in identifying and extracting about 68% of the causal relations that were clearly expressed within a sentence or between adjacent sentences in *Wall Street Journal* text. Of the instances that the computer program identified as causal relations, 72% can be considered to be correct.

The automatic method was used in an experimental information retrieval system to identify causal relations in a database of full-text *Wall Street Journal* documents. Causal relation matching was found to yield a small but significant improvement in retrieval results when the weights used for combining the scores from different types of matching were customized for each query -- as in an SDI or routing queries situation. The best results were obtained when causal relation matching was combined with word proximity matching (matching pairs of causally related words in the query with pairs of words that co-occur within document sentences). An analysis using manually identified causal relations indicate that bigger retrieval improvements can be expected with more accurate identification of causal relations. The best kind of causal relation matching was found to be one in which one member of the causal relation (either the cause or the effect) was represented as a wildcard that could match with any term.

The study also investigated whether using *Roget's International Thesaurus* (3rd ed.) to expand query terms with synonymous and related terms would improve retrieval effectiveness. Using Roget category codes *in addition to* keywords did give better retrieval results. However, the Roget codes were better at identifying the non-relevant documents than the relevant ones.

AUTOMATIC IDENTIFICATION OF CAUSAL RELATIONS IN TEXT AND THEIR USE FOR IMPROVING PRECISION IN INFORMATION RETRIEVAL

by

CHRISTOPHER SOO-GUAN KHOO
B.A., Harvard University
M.S., University of Illinois at Urbana-Champaign

DISSERTATION

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Information Transfer
in the School of Information Studies of Syracuse University

December, 1995

Approved ____________________
Professor Robert N. Oddy

Date ____________________

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My heartfelt thanks to my two advisors, Prof. Sung Myaeng and Prof. Bob Oddy. Prof. Myaeng was my dissertation advisor before he left Syracuse Unversity to take up a position in his native country of South Korea. I was fortunate to have Prof. Oddy take over supervision of my dissertation. I have benefited greatly from their advice, encouragement and wisdom -- not just in the course of this research but also throughout my years in the doctoral program.

I am grateful to my committee members, Professors Jaklin Kornfilt, Barbara Kwasnik, Susan Bonzi, Murali Venkatesh and Noriko Kando, for taking time to read and comment on earlier versions of the dissertation, and for their support and advice at critical moments in the course of my doctoral studies.

I would like to thank

- Harper Collins Publishers for use of the machine-readable file of the *Roget's International Thesaurus* (3rd ed.)
- Longman Group UK Limited for use of the machine-readable file of the *Longman Dictionary of Contemporary English* (new ed.)
- *Dow Jones & Co.* for use of the *Wall Street Journal* text (obtained through the *National Institute of Standards and Technology*)
- the *National Institute of Standards and Technology* for use of the Tipster/TREC information retrieval test collection
- *BBN Systems and Technologies* for use of the POST part-of-speech tagger
- Chris Buckley for use of the evaluation program from the SMART system
- Prof. Sung Myaeng and Prof. Elizabeth Liddy for use of some computer programs and processed text from the DR-LINK project.

This dissertation would not have been successfully completed without the help, support and prayers of many individuals.

My deepest gratitude to my wife and helpmeet, Soon-Kah Liew, for her patience and love through these many years, and for her help with some parts of this research.

I owe much to my family and my wife's family: my parents, Mr and Mrs Khoo Heng Chuan, for their encouragement and unceasing prayers; my mother-in-law, Mdm Lim Chui Hiong, for all the effort and energy she has expended in taking care of our affairs in Singapore and for taking care of our financial needs; my siblings and my wife's siblings for their concern, prayers and financial support.

The members of the international Bible study group at Slocum Heights gave us many years of fellowship and prayer support. Our grateful thanks also to many Christians who have remembered us in their prayers.

Finally, I praise and glorify God for the many blessings we have received and for sustaining us through this difficult journey.

CHAPTER 1
INTRODUCTION

1.1. Overview

This study has two parts. In the first part of the study, an automatic method for identifying causal relations in natural language English text was developed. The second part of the study investigated whether the effectiveness of an information retrieval system could be improved by matching causal relations expressed in documents with the causal relations specified in the user's query statement, and using the information from causal relation matching in predicting whether a document was likely to be relevant to the user. The automatic method for identifying causal relations developed in the first part of the study was used in the second part of the study to identify causal relations in the test database.

The automatic method for identifying causal relations developed in this study makes use of a set of linguistic patterns, each pattern being a sequence of words and syntactic categories that usually indicates the presence of a causal relation. A computer program was written to use the set of patterns to detect causal relations expressed in text, and to extract from the text the *cause* and the *effect*. No inferencing from commonsense knowledge or domain knowledge was used in this study. The computer program uses only linguistic clues to identify causal relations. The goal was to develop a method for identifying causal relations that was appropriate for use in an information retrieval system which catered to a heterogeneous user population with a wide range of subject interests. The set of linguistic patterns was constructed based on an extensive literature review (reported in Chapter 3) and then refined by repeatedly applying the patterns to a sample of sentences, modifying the patterns to eliminate the errors, and applying the patterns to a new sample of sentences. It is of interest to find out how effective a program that uses linguistic patterns can be in the task of identifying causal relations in text:

> **Research question 1**
> How effectively can cause-effect information expressed in sentences be identified and extracted using linguistic patterns?

The computer program and the linguistic patterns for identifying causal relations were used in several retrieval experiments to investigate the second research question relating to the use of causal relation matching in information retrieval:

> **Research question 2**:
> Can the information obtained by matching causal relations expressed in documents with causal relations expressed in the user's query statement be used to improve retrieval effectiveness over just matching terms without relations?

In the retrieval experiments, the computer program was used to identify causal relations in documents. The terms and the causal relations identified in the documents were then matched with the terms and causal relations in the user's query statement. The purpose of the study was to find out whether it improved retrieval results to calculate the similarity between the query and the documents *using the additional information of causal relation matches*. This study thus represents one attempt to go beyond term matching in information retrieval, and to see if taking into account relations between terms, in this case *causal relations*, can help improve retrieval effectiveness.

An interesting feature of this study was the use of the *Roget's International Thesaurus* (3rd ed.) for conflating synonyms and related words. Each noun, verb, adjective and adverb in the documents and query statements was assigned one or more Roget category codes by the retrieval system. The use of Roget category codes can be seen as a method of query expansion: words in the query statements were, in effect, expanded with other words that had the same Roget category code.

Relation matching, in particular *causal relation matching*, is mainly a *precision-enhancement* device. Relation matching allows the retrieval system to increase the precision of retrieval by using additional "relational" information to identify a smaller subset of documents that are more likely to be relevant from among those documents retrieved using term matching. It may be helpful to complement this precision-enhancement procedure with a *recall-enhancement* procedure. The use of Roget category codes is a *recall-enhancement* procedure in that it makes it possible to retrieve some documents that are not retrievable using just the words given in the query statement.

In the next section, I discuss the motivation for focusing on causal relations. I then attempt to define what the *causal relation* is. Finally, I specify the kind of information retrieval system that was assumed in this study.

1.2. Motivation for This Study

This study represents one attempt at going beyond using just keyword matches in estimating how likely a document is relevant to a user's information need. The use of keyword matching for information retrieval has been well explored, and effective methods for using keyword matching are well-known (Salton & McGill, 1983). It does not seem likely that substantial improvement in retrieval effectiveness can be obtained with further research in keyword matching methods. Yet, the retrieval performance achievable today is still not very good (Harman, 1995). There is a lot of room for improvement.

Two approaches for improving retrieval effectiveness that are actively being explored are the use of automatic query expansion (i.e. expanding query words with related words) and relation matching. The former approach focuses on recall enhancement, the latter on precision enhancement. This study focused on relation matching, in particular, *causal relation* matching, although query expansion using Roget category codes was also performed.

Results of previous studies using relation matching (reviewed in Chapter 2) have been disappointing. Some retrieval improvement from relation matching was obtained in some studies but the improvements were not greater than is obtainable just by using term proximity matching, i.e. by assuming that the desired relation between two query terms is present in the document just from the fact that they occur within a document sentence or in close proximity, and giving the document a higher retrieval score when this occurs.

There are two factors that affect the usefulness of relation matching. The first factor is the *accuracy factor*. It is difficult to identify relations in text accurately using automatic means. There is usually a substantial error rate associated with the automatic identification of relations in text using present day text processing computer programs.

The second factor is what I call the *relational ambiguity factor*. Relation matching will give better retrieval results than term proximity matching to the extent that it is difficult to predict the relation between two terms just from the fact that they occur near to each other in a document. For example, if the words *eat* and *apple* occur in the same sentence, we can quite confidently predict that *apple* has the "patient" (i.e. object) relation to the word *eat* without even reading the sentence. There is little relational ambiguity in this case. If a query statement contains the relation

eat ->(patient)-> apple

using sophisticated natural language processing to identify this relation in document sentences will probably not yield better retrieval results than just searching

eat (same sentence) apple

i.e. searching for documents where *eat* and *apple* occur within the same sentence. The greater the relational ambiguity between two query terms occurring in the document, the more likely will relation matching help improve retrieval results. The degree of relational ambiguity between two terms increases with the distance between the terms in the text. If the words *eat* and *apple* are adjacent in the text, there is no doubt what the relation between the two terms is. However, if the two terms occur further apart, e.g. in adjacent sentences, then there is more doubt about what relation, if any, exists between the two terms. There is more relational ambiguity in this case. I shall discuss one implication of this later in the section.

To reduce the problem of accuracy in identifying relations in text, I felt that researchers should study one type of relation at a time, instead of trying to handle many relations. By looking at just one type of relation, one can

1. focus one's effort at developing a method to identify one type of relation as accurately as possible
2. investigate the usefulness of that type of relation for information retrieval.

Some types of relations may be more useful than others in improving retrieval effectiveness. In this study, I chose to focus on the causal relation.

Why causal relation? Previous studies of relation matching have used both syntactic and semantic relations[1]. Semantic relations are preferable because the same semantic relation can be expressed in many syntactic forms. In matching semantic relations, we are performing matching across different syntactic relations. Syntactic relation matching may yield fewer matches than semantic relation matching. Also, the user judges whether a document is relevant or not based on meaning (e.g. semantic relations) rather than form (syntactic relations).

Semantic relations can be at different levels of abstraction. A high level relation is one that can be decomposed into more primitive concepts and relations. Consider the following conceptual graph (a type of semantic representation):

[person:John] ->(eat)-> [apple] [2]

The conceptual graph expresses the fact that John has an "eating" relationship with an apple. The relation *eat* is a high-level relation that can be decomposed into the concept *eat* and the "case relations" (also referred to as thematic relations or theta roles) *agent* and *patient*. The above conceptual graph thus expresses the same thing as the following conceptual graph:

[person:John] <-(agent)<- [eat] ->(patient)-> [apple]

Case relations are low-level semantic relations that exist between the main verb of a clause and the other constituents of the clause (Fillmore, 1968; Somers, 1987). Two recent studies, Lu (1990) and

[1] By *syntactic relation,* I mean the relation between two terms derived from the syntactic structure of the sentence. By *semantic relation,* I mean the logical or conceptual relation expressed in the text but not wholly dependent on the particular syntactic structure of the sentence.

[2] A word within square brackets is a label for a concept. A word within round brackets is a label for a relation. Arrows indicate the direction of a relation.

Myaeng and Liddy (1993), did not obtain good retrieval results with case relation matching. Case relations exist between terms that occur very close together in a sentence -- usually in adjacent positions and always within the same clause. With terms that occur so close together, relational ambiguity is less likely. Higher-level relations can exist between terms that occur further apart in the document. Relational ambiguity is more likely when terms are further apart. Relation matching is thus more likely to be helpful with higher-level relations than with case relations. From this perspective, the *causal relation* looks like a good candidate for use in information retrieval. The causal relation can exist between two terms within a clause, between two clauses, or between two sentences.

The *causal relation* is an important relation in most fields of study. The purpose of most social science research is to find causal relations, i.e. to find out what factor tends to cause which phenomenon. Of the first 150 query statements in the TIPSTER/TREC information retrieval test collection (Harman, 1994a)[3], about half of them contain one or more causal relations. So, the causal relation is quite prevalent in the queries sampled by the organizers of the TIPSTER/TREC project. It may also be prevalent in the queries of other user populations.

Finally, the method for identifying causal relations developed in this study represents a contribution to the area of knowledge extraction -- the development of automatic methods for extracting knowledge from natural language text. Knowledge extraction research has important implications for knowledge-based systems. The "knowledge-acquisition bottleneck" has been identified as a central problem in artificial intelligence and knowledge-based systems research (Chan, 1995; Feigenbaum, 1984; Gaines & Shaw, 1991; Marik & Vlcek, 1992; Tafti, 1992). Currently, knowledge for expert systems is acquired mainly through interviews with human experts and by manually extracting the knowledge from textbooks. The knowledge acquisition process is thus very slow. Automatic extraction of knowledge from text offers some hope of speeding up the process (Yuan, Chang & Suen, 1994). Causal knowledge is an important kind of knowledge used in expert systems, especially in model-based or "deep" expert systems that attempt to reason from first principles or from a model of the domain rather than rely on heuristics or rules of thumb (Chorafas, 1990; Jackson, 1989 and 1990; Kaplan & Berry-Rogghe, 1991; Selfridge, 1989).

1.3. What is a *Causal Relation*?

In this section I attempt to characterize what we mean when we say that *A caused B* or that *an event of the kind A causes an event of the kind B*? I'm concerned only with the layman's commonsense idea of causation, and not causation as defined in any particular field of study. I shall first review philosophical arguments about people's concept of causation. I shall then review experimental evidence concerning how people ascribe *cause* and infer the strength of a causal relation. Knowing how people ascribe cause and infer the strength of a causal relation gives us a fuller understanding of the concept of causation.

1.3.1. Causation in philosophy

Causation is a complex concept. Philosophers have grappled with the concept for centuries. Books are still being written on the subject (e.g., Emmet, 1985; Fales, 1990; Owens, 1992; Sosa & Tooley, 1993; Strawson, 1989; Tooley, 1987). Short surveys of the issues are given in White (1990) and in the introduction to Sosa and Tooley (1993). Two philosophers who have contributed a great deal to our understanding of the concept of causation are David Hume and John Stuart Mill.

[3] This test collection is described in Chapter 5 Section 5.3.1.

For Hume (1740/1965), causation comprises the following three conditions: 1. *contiguity* in time and place, 2. *priority* in time, and 3. *constant conjunction* between the cause and the effect. When a person finds, from experience, that an event of the kind A is always followed by an event of the kind B, the person comes to conclude that event A causes event B. For Hume, causation is nothing more than the association in the mind of two ideas (e.g. event A and event B) as a result of experiencing their regular conjunction.

J.S. Mill (1872/1973) argued that constant conjunction is not sufficient for inferring causation, unless the conjunction is also unconditional. For example, although day invariably follows night, we do not infer that night causes day. Day follows night only provided the sun continues to rise.

Mill (1872/1973) described four methods by which one can determine that A causes B: the method of agreement, the method of difference, the method of residues, and the method of concomitant variations. Perhaps the most important is *the method of difference* which involves comparing two instances, one in which an event of the kind A occurs, and one in which A does not occur. If the two instances are similar in every respect except that for the instance in which A occurs, it is followed by B, but for the instance in which A does not occur, B also does not occur, then we can conclude that A causes B. Similarly, if we find two instances, one in which B occurs and one in which B does not occur, and we find that when B occurs, it is preceded by A, but when B does not occur, A also does not, then we can also conclude that A causes B.

Mackie (1980) argued that the layman does use this kind of reasoning when deciding whether there is a causal relation. In deciding whether a particular event A caused an event B, we engage in the counterfactual (i.e. contrary-to-fact) reasoning that involves asking whether B would have occurred if A had not occurred. If B would not have occurred had A not occurred, we conclude that A caused B.

Mill's (1872/1973) *method of difference* has been extended to distinguish between *necessary* and *sufficient* causes. If when an event of the kind A occurs, an event of the kind B always follows, but when A does not occur, B sometimes occurs and sometimes not, then A is a *sufficient* though not a *necessary* condition for B to occur. On the other hand, if when A does not occur, B never occurs, but when A occurs, B sometimes occurs and sometimes not, then A is a *necessary* though not a *sufficient* condition for B to occur.

Mill (1872/1973) also contributed to our understanding of causation by pointing out that an effect is usually the result of a conjunction of several causes, even though in practice one of these causes is singled out as *the cause*:

> It is seldom, if ever, between a consequent and a single antecedent, that this invariable sequence subsists. It is usually between a consequent and the sum of several antecedents; the occurrence of all of them being requisite to produce, that is, to be certain of being followed by, the consequent. In such cases it is very common to single out one only of the antecedents under the denomination of Cause, calling the others merely Conditions. . . . The real Cause, is the whole of these antecedents; and we have, philosophically speaking, no right to give the name of cause to one of them, exclusively of the others. . . . If we do not, when aiming at accuracy, enumerate all the conditions, it is only because some of them will in most cases be *understood without being expressed* [my emphasis], or because for the purpose in view they may without detriment be overlooked. (Mill, 1872/1973, pp. 327-329).

Mackie (1980) referred to the background conditions that are understood without being expressed as *the causal field*.

Mackie (1980) pointed out that under different circumstances, different antecedents will be selected as *the cause*. How do people decide in a particular situation which antecedent to select as *the cause*? Hesslow (1988) listed ten criteria, including unexpected conditions (other conditions being understood without being expressed), precipitating causes, abnormal or unusual conditions, deviation from a theoretical ideal, responsibility (causal statements may have an evaluative component), predictive value, and the subjective interest of the observer.

Hesslow (1988) suggested that which criterion is used to select *the cause* depends on the *contrast event*. According to Hesslow, when we ask why event B occurred, we are implicitly asking why event B occurred in comparison to some other event, the *contrast event*. For example, when we ask why the barn caught fire, we are unconsciously comparing the barn with our conception of a normal barn. Since the normal barn has not caught fire, an *explanatorily relevant* cause for the barn's catching fire must be some abnormal condition. If we were comparing the barn that caught fire with the same barn before it caught fire, we might select the precipitating or immediate cause instead. So, for Hesslow, a cause explains a certain event in comparison to a, possibly implicit, contrast event.

The argument so far leads us to characterize the *cause* as some event or fact that explains why something occurred rather than something else, under the circumstances (the assumed causal field). However, we have not established whether the *cause* is a necessary condition, a sufficient condition or both. Mackie (1980) argued that *cause* is a necessary condition but not necessarily a sufficient condition. It is a sufficient condition only *in a weak sense*. For example, *sexual intercourse* is sometimes, but not always, followed by *pregnancy*; and without *intercourse*, *pregnancy* does not occur (disregarding recent developments like artificial insemination).[4] Clearly *intercourse* is not a sufficient condition for *pregnancy*, although it is a necessary condition. Suppose there was a particular instance in which *intercourse* was followed by *pregnancy*. We would say that for that particular instance the *intercourse* caused the *pregnancy*. For that instance, *intercourse* was a sufficient condition for *pregnancy* to occur. That was what Mackie meant by *a weak sense* of sufficient condition. Mackie characterized *cause* as a necessary condition under the circumstance, and a sufficient one in a weak sense.

Jaspars, Hewstone and Fincham (1983) and Jaspars (1983) found evidence that whether a cause is a necessary, sufficient, or necessary and sufficient condition varies with the type of entity being considered for causal status. Cause is likely to be attributed to a person if the person is a sufficient condition. Necessity does not appear to be important when a person is a candidate for causal status. On the other hand, cause is likely to be attributed to the circumstances or situation if the situation is a necessary condition. Sufficiency is not so important for situational causes. Cause is ascribed to a stimulus when it is both a necessary and sufficient condition. So, "a personal cause is seen more as a sufficient condition, whereas situational causes are conceived primarily as necessary conditions." (Jaspars, Hewstone & Fincham, 1983, pp. 16-17)

Mackie (1980) also pointed out that our concept of causation includes some presumption of a continuity from the cause to the effect, a mechanism by which the cause generates the effect. We conceive of the effect as being "fixed" by the cause. In the words of Owens (1992), "a cause is an event which ensures that its effects are no coincidence." (p. 23)

There has been some debate in the literature about whether a cause or effect can be a *fact*, or whether it has to be an *event* or implied event. Many writers (e.g., Davidson, 1980; Kovalyova, 1988) contend that only events can be linked by a causal relation. Nouns that are expressed as causes in sentences are actually ellipses for implied events. However, Mackie (1980) argued that only certain facts or states in a causal event are actually *causally relevant* to the effect. For example, when we crack a nut by hitting it with a hammer, we can say that *the hammer hitting the nut* is the event that caused the nut to crack. However, it is *the fact that* the force of the hammer

[4] This example is from Einhorn and Hogarth (1986).

exceeded a certain amount that caused the nut to crack. So, a cause can be a *fact*. However, Peterson (1981) argued that "the force of the hammer exceeding a certain amount" is not *the cause* of the nut cracking, but *the reason* it cracked. He distinguished between *cause* and *reason* (or causal explanation). In this study, I do not distinguish between *cause* and *reason*.

Newman (1988) argued that causes and effects can only be objects and properties. For Newman, events are not "real." Only objects and properties are real. Events are "arbitrary objects" which have to be constructed out of real objects and properties. The only way to deal with events is via their components, i.e. objects and their properties:

> Unlike a unified object, an event or a set is an entity only because someone insists on regarding it as such. It is possible to say that events "exist", in the same way that it is possible to say that sets "exist", provided it is not thereby implied that events *as such* [Newman's emphasis] are real. Events do not exist in the way unified objects exist. (p. 548)

In this study, I do not limit *cause* to particular types of entities. For the purpose of this study, a cause can be an event, a fact, an object, a state or property of an object, or an agent.

A related issue is whether in the linguistic "deep structure", a cause or effect can be a noun, or whether it has to be a proposition or predicate (Owens, 1992, pp. 60-62). Dakin (1970) argued that noun phrases linked by causal relations are acting as pro-sentences, i.e. they represent a sentence or clause in the deep structure. As this study deals only with the surface structure of sentences, the text processing used in this study assumes that causes and effects can be in the form of noun phrases or clauses.

In the foregoing, I have reviewed some literature on causation that mostly uses philosophical arguments and plausible examples to analyze the concept of causation. There is a body of research on attribution theory (a branch of social psychology) that provides experimental data about how people decide whether *A* caused *B*.

1.3.2. Experimental results from attribution research

Attribution theory deals with how people ascribe cause or responsibility[5] for an event to one or more elements of a situation. Research in attribution theory typically involves presenting verbal vignettes describing certain events (the target events) to subjects. The vignettes typically involve a person, a stimulus and an event, as in the following example from McArthur (1972):

> John laughs at the comedian.

[5] Cause and responsibility are often used interchangeably in the literature on attribution theory. However, Shultz & Schleifer (1983) argued that causation refers essentially to event generation and responsibility to moral evaluation of an actor.

John is the *person*, the comedian is the *stimulus*, and the event happens at a particular occasion (the *circumstances*). Given a vignette, the subjects have to decide whether the cause of the event is the person, the stimulus, the circumstances or some conjunction of the three elements. To help them decide, subjects are typically given the following three kinds of information: 1. *Consensus* information, describing whether the behavior generalizes over other persons; 2. *Distinctiveness* information, describing whether the behavior generalizes over other stimuli; and 3. *Consistency* information, describing whether the behavior generalizes over other occasions. The following is an example of consensus, distinctiveness, and consistency information for the target event "John laughs at the comedian":

> Hardly anyone/almost everyone laughs at the comedian. (low/high consensus)
> John laughs at hardly any/almost every other comedian. (high/low distinctiveness)
> In the past, John has hardly ever/almost always laughed at this particular comedian. (low/high consistency)

Jaspars, Hewstone and Fincham (1983) proposed the *inductive logic* model of how people use consensus, distinctiveness, and consistency information to attribute cause. The inductive model is based on the idea that people can infer the *necessary* condition(s) for the event from consensus, distinctiveness and consistency information. For example, given the information "Hardly anyone laughs at the comedian" (low consensus), people can infer that the person John is a necessary condition for the event since other people do not laugh at the comedian. The necessary condition(s) are jointly sufficient for the event, and are considered to be the cause of the event. Put it another way, the consensus, distinctiveness and consistency information provides information for people to use in counterfactual reasoning. Consensus information allows people to answer the counterfactual question "if John hadn't been present, would the event of *laughing at the comedian* occurred?"

In experiments by Jaspars (1983) and Hilton and Jaspars (1987), the inductive logic model was shown to predict subjects' responses fairly well, except when the subjects were given high consensus, low distinctiveness, and high consistency information. The inductive logic model predicts no attribution at all for this configuration of information since neither the person nor the stimulus nor the circumstances can be inferred to be a necessary condition under this configuration. Subjects, however, made 70% of their attributions in this case to the person, the stimuli or both.

The *abnormal conditions focus model* of Hilton and Slugoski (1986; also Hilton, 1988) makes the same predictions as the inductive logic model, except in the case of the problematic situation of high consensus, low distinctiveness, and high consistency configuration where it explains subjects' responses by assuming that they make use of their general knowledge about what the statistically normal case is. This statistically normal case is used as the *contrast case*. The *abnormal conditions focus model* assumes that consensus, distinctiveness and consistency information serves to define contrast cases that highlight certain aspects of the target event as abnormal. These abnormal aspects are then considered to be the cause of the event. For example, low consensus information ("Hardly anyone laughs at the comedian") throws the target person, John, into focus as abnormal. This approach is consistent with Hesslow's suggestion, mentioned earlier, that a cause explains an event *in comparison to a contrast event*.

1.3.3. Causation as covariation

Consensus, distinctiveness and consistency information typically used in attribution research present extreme conditions. For example, the subject is told either that almost everyone laughs at the comedian or that hardly anyone laughs at the comedian. In "real life," it is more likely the case that *some* people laugh at the comedian. Sometimes quantitative information is available: 30 percent of the people laugh at the comedian. Research has found that people are generally capable of making use of such covariation[6] information to assign cause and to judge the strength of the causal relation. Kelley (1973) expressed the covariation principle as follows:

> An effect is attributed to the one of its possible causes with which, over time, it covaries. (p. 108)

This idea can be traced back to Mill's (1872/1973) *method of concomitant variations*, and Hume's (1738/1911, Bk. I, Pt. III) idea of *frequent conjunction*.

Kelley (1967, 1973) suggested that people perform some kind of analysis of variance (ANOVA) to infer the cause of an event. For the example given in the last section of John laughing at the comedian, Kelley would say that people perform a naive form of ANOVA involving three independent variables, persons, stimuli and circumstances, in a factorial design. Assuming that each independent variable has two levels (i.e. the *persons* variable has two levels, *John* and *other people*; the *stimuli* variable has two levels, *this comedian* and *other comedians*; and the *circumstances* variable has two levels, *this occasion* and *other occasions*), then the ANOVA model forms a cube with eight cells. This has been referred to in the literature as Kelley's cube. Studies by McArthur (1972) and Orvis, Cunningham, and Kelley (1975) provide partial support for this model.

There is a number of problems with Kelley's cube as a model of human causal induction. The dependent variable has discrete values (e.g., either the event of *laughing* occurs or it does not occur), and so ANOVA is not really an appropriate statistical model to use. Moreover, the cell in the ANOVA model representing the conjunction of John, this comedian and this occasion has a sample size of one. Furthermore, consensus, distinctiveness and consistency information typically used in attribution research provide information only for four cells in Kelley's cube. Information such as whether other people laugh at other comedians on other occasions is usually not used in attribution research. Nevertheless, Kelley's ANOVA model has been very influential in attribution research and has inspired several reformulations of the model.

Since Kelley (1967), three important models of how people use covariation information to infer a causal relation have been proposed: the likelihood ratio model (Ajzen & Fishbein, 1975), the linear combination model (Downing, Sternberg & Ross, 1985; Schustack & Sternberg, 1981), and the probabilistic contrast model (Cheng & Novick, 1990 and 1992).

Several studies found certain biases in people's use of covariation information for inferring the cause or for predicting future events. For example, some studies found that subjects did not make adequate use of base rates or *consensus* information (e.g., Hansen & Donoghue, 1977; McArthur, 1972; Nisbett & Borgida, 1975; Schustack & Sternberg, 1981). On the other hand, there are other studies that found no biases under certain conditions (e.g., Hewstone & Jaspars 1983; Kulik & Taylor, 1980). Ajzen and Fishbein (1975) said that "whenever apparent attribution biases have been consistently obtained, the results could also be interpreted in terms of a more rational information processing model." (p. 275)

[6] Covariation between two events may be defined in terms of their co-occurrence, i.e. the degree to which one event occurs more often in the presence than in the absence of the other event (Alloy & Tabachnik, 1984).

Kassin (1979) in his review of the literature on the use of consensus information concluded that base rates provided by the experimenter are underutilized by the subjects when the base rates are either redundant or highly inconsistent with the subjects' implicit or preconceived base rates. However, increased use of base rates occurs when the base rates are made more salient or meaningful, for example by presenting the base rate information just prior to the judgment (Ruble & Feldman, 1976), or by informing the subjects that the base rate was obtained using random, representative sampling (Wells & Harvey, 1977).

Ajzen (1977) and Tversky and Kahneman (1980) found that subjects do make appropriate use of the base rate when the base rate has a causal implication. In one study by Tversky and Kahneman (1980), subjects were asked to estimate the probability that the cab involved in an accident was *blue* rather than *green*. One group of subjects were given the base rate that 85 percent of the cabs in the city were *green*. Another group of subjects were given the base rate that 85 percent of cab accidents involved *green* cabs. The first group of subjects did not make use of the base rate whereas the second group did. In the second case, the base rate that 85 percent of cab accidents involved green cabs carries the implication that green cab drivers were less careful drivers.

Covariation is not the only information that people use to infer causation. Theories and prior beliefs also play an important part. In an extensive review of the literature on the ability of people (and animals) to detect and use covariation information, Alloy and Tabachnik (1984) concluded that people can assess covariation fairly accurately and use the information to make statistically-based causal attributions, provided that there is no prior expectation or belief about the covariation or causal explanation. Prior expectation will bias the assessment of covariation or causal attribution in favor of the prior expectation, and the degree of bias depends on the relative strengths of the expectation and the "objective" covariation information.

In their review of the literature, Einhorn and Hogarth (1986) listed the following "cues to causality" (i.e. information people use to infer a causal relation):

- covariation of the suspected cause, *X*, with the effect, *Y*
- the degree to which *X* is a difference in the background (e.g., an abnormal condition is more likely to be seen as cause)
- the temporal order of *X* and *Y*
- contiguity of *X* and *Y* in time and space
- similarity of *X* and *Y*, including physical similarity and congruity of duration and magnitude
- the causal chain strength (the extent to which a plausible theory or scenario can be constructed to link *X* to *Y*)
- the lack of plausible alternative causes (a causal explanation is *discounted* to the extent that there are plausible alternatives.)

I would like to highlight the importance of *theory* in causal inference. Covariation of X and Y, however strong, does not lead one to infer a causal relation if it is not plausible according to some theory. Shultz (1982) demonstrated that covariation information is not used when it conflicts with knowledge of how simple physical effects are produced. Said Einhorn and Hogarth (1986, p. 10), "causal relations must be *achieved* in the sense that prior knowledge and imagination are needed to construct a schema, scenario, or chain, to link cause and effect."

1.3.4. So what is causation?

The concept of causation is surprisingly complex. Research into what we mean by causation and how we ascribe cause is still active. I have outlined the main features of the concept of causation. A brief definition of causation is that *A* causes *B* if *A* in some sense generates or fixes the occurrence of *B* in the circumstance (the assumed causal field). Whether the cause is a necessary, sufficient or both a necessary and sufficient condition depends on what *A* is. People are capable of using covariation information in causal attribution, but the use of covariation information is tempered by the theories and prior beliefs that people have.

People ascribe cause to a wide range of phenomena. Besides physical causation, cause can refer to reason (for an action), motivation, psychological causation, human actions and interactions, statistical laws, teleological cause[7], etc. In this study, I take *causation* to include all the various types of causation.

1.4. What is an Information Retrieval System?

There are many types of information retrieval systems. They range from simple online library catalogs to full-text document retrieval systems, and to systems that retrieve document paragraphs in response to a query. It can even be argued that question-answering systems are information retrieval systems.

In this study, an information retrieval system was assumed to be a full-text document retrieval system that retrieves documents in response to the user's information need statement (henceforth, *query statement*). The retrieval system attempts to rank the documents in the database according to the likelihood of relevance or, alternatively, the degree of relevance to the user's information need.

Relevance can be defined in various ways (Schamber, Eisenberg & Nilan, 1990). There are two main approaches:

- relevance as *conceptual relatedness*. This "refers to a fundamental connection or fit between concepts in information need and concepts in information, or some idea of what information is about beyond simple topicality." (Schamber, Eisenberg & Nilan, 1990)
- relevance as *usefulness*. This refers to whether the retrieved document is useful to the user in some way.

The definition of *relevance* assumed in this study was *conceptual relatedness*. The query statements used in this study specify what information a document must contain in order to be relevant. The query statements do not state the users' purpose for doing the information search or what use the users are going to make of the information obtained. Furthermore, the relevance judgments were not done by the same people who formulated the queries, but by judges who made the relevance judgments based on the query statements.

This study dealt only with *subject searching*, as opposed to *specific-item* (or *known-item*) searching. Specific-item searching refers to searching by author and/or title, while subject searching is generally taken to mean searching by topic or subject content.

The test collection used in this study was a subset of the TIPSTER/TREC test collection used by the participants of TREC-1 and TREC-2 conferences (Harman, 1993a; Harman, 1993b;

[7] A *teleological cause* is the distant effect or goal towards which an event is directed.

Harman, 1994a).[8] The document collection used comprised approximately five years of the full-text of the *Wall Street Journal*. The *Wall Street Journal* documents were selected for this study because I had worked with them in the DR-LINK project (Liddy & Myaeng, 1994; Myaeng & Liddy, 1993; Myaeng, Khoo & Li, 1994), and so was familiar with it. Also, the documents had been processed in several ways in the DR-LINK project (described in Chapter 4 Section 4.3 and in Chapter 5 Section 5.3.1) that made them more convenient to use.

1.5. Summary

This study explored the use of causal relation matching for improving the effectiveness of information retrieval systems. An automatic method was developed for identifying causal relations in text using linguistic patterns. The use of Roget category codes for query expansion was also investigated.

I have argued that research on relation matching in information retrieval should be carried out by focusing on one type of relation at a time so that the automatic identification of the relation can be made as accurate as possible and we can investigate the usefulness of each type of relation in improving the retrieval results. I have also argued that higher-level relations are more likely to be useful than low-level relations like case relations. The causal relation appears to be a good relation to investigate because causal knowledge is an important kind of knowledge and finding causal relations between things is important to many fields of study.

The causal relation is a surprisingly complex relation. I have suggested that when we say *A* causes *B*, we mean that *A* in some sense generates or fixes the occurrence of *B* in the circumstance. Whether the cause is a necessary, a sufficient or both a necessary and sufficient condition depends on what *A* is and what the circumstance is.

This study assumed an information retrieval system to be a full-text document retrieval system. The definition of *relevance* assumed in this study was that of *conceptual relatedness* rather than *usefulness to the user*.

[8] There has since been a third and a fourth TREC conference with more query statements, documents and relevance judgments added to the test collection (Harman, 1995).

CHAPTER 2
LITERATURE REVIEW

2.1. Introduction

This literature review is divided into two parts. The first part surveys the literature on the automatic extraction of causal knowledge from text. The second part reviews previous research on the use of relation matching in information retrieval.

2.2. Automatic Identification of Causal Relations in Text

Studies on the automatic identification of causal relations in text and the automatic extraction of causal knowledge from text have focused on the following kinds of text:

1. episodic text (also called narrative text), which describes a series of related events involving human actions (e.g., a story)
2. short explanatory messages that a human user might enter into a computer system as part of a human-computer dialog on a particular subject
3. expository text of the kind found in textbooks.

Researchers have used somewhat different approaches for each of the three types of text. However, most of the studies have focused on the use of knowledge-based inferencing to identify causal relations in text.

Research on the automatic identification of causal relations in *episodic text* is usually undertaken with the broader goal of developing computer programs that can "understand" stories, and perform comprehension tasks like answering questions about stories and summarizing the stories. Numerous studies have shown that inferring causal relations between sentences is an important part of text comprehension and story understanding (e.g., Fletcher & Bloom, 1988; Keenan, Baillet & Brown, 1984; O'Brien & Myers, 1987; Trabasso, 1989; Trabasso, Secco & van den Broek, 1984; van den Broek, 1989).

In the studies on episodic text, the inferencing of causal relations is usually done using:

- *scripts* or *schemas* that encode extensive knowledge (including causal knowledge) that people have about particular types of episodes (Schank, 1982; Schank & Abelson, 1977)
- *plan-based inferencing*, using knowledge about the *goals* that people have in certain situations and the possible *plans* that can satisfy the goals (Wilensky, 1983).

Examples of studies on episodic text are Bozsahin and Findler (1992), Cullingford (1978), Lebowitz (1980), Mooney (1990), Schubert and Hwang (1989), and Wilensky (1978). These studies sought to find out what kind of knowledge and what kind of inferencing are needed to accurately infer causal relations between events described in the text and to infer events that are implied in the text.

Some of the studies on episodic text (e.g., Bozsahin & Findler, 1992; Pazzani, Dyer & Flowers, 1987) have developed techniques for extracting from a set of episodic texts knowledge about what event tends to cause what other event in a particular context. The causal knowledge is obtained by comparing all the episodes stored in the system's memory, and generalizing from the similarities found in these episodes. Causal relations are inferred using the same kind of heuristics that people use to infer causation. For example, the study by Bozsahin and Findler (1992) used Hume's (1740/1965) rule of *frequent conjunction*: if an event of the type A is followed by an event of the type B in many of the episodes stored in the system's memory, then infer that event A tends to cause event

B. Besides general heuristics like this, the system may also use domain-specific heuristics to infer causal relations.

Studies that deal with episodic text typically make little use of linguistic clues to identify causal relations. Presumably, explicit linguistic indications of cause and effect, such as *because*, *if ... then*, and *as a result of this*, do not occur often in episodic text.

A second group of studies have focused on identifying causal relations in short explanatory messages of the kind that a human domain-expert might enter into the knowledge acquisition component of an expert system. The capability of extracting causal knowledge from natural language explanations will make it easier for an expert system to acquire new knowledge of a domain directly from a human expert. When there is ambiguity about whether a causal relation between two events is expressed in the text, the system can attempt to resolve the ambiguity by using the knowledge or model of the domain that it already possesses to check whether a causal relation between the events is possible. Selfridge (1989) has reviewed the main issues involved in the automatic acquisition of causal knowledge from human experts.

I have found only two studies that attempted to develop computer programs for extracting causal knowledge from short explanatory messages: the study by Selfridge, Daniell and Simmons (1985) and that by Joskowsicz, Ksiezyk and Grishman (1989). Both studies used a physical system as the domain, and both had the goal of building a model-based expert system for the purpose of fault diagnosis (i.e. diagnosing the cause of equipment failure).

Selfridge, Daniell and Simmons (1985) developed their system, called CMACS, to acquire knowledge about how a gasoline engine works. The system was first programmed with basic knowledge about the components of the engine and their possible states and behavior. The system then processed a short text about gasoline engines to obtain information about causal relations between the different states of the engine components. The causal relations extracted from the text were used to build a computer model of the engine. The system then checked the model for incomplete knowledge (e.g., component states not linked by a causal relation) and asked the human expert for an explanation of the unexplained events. The linguistic processing and inferencing used to identify causal relations in text are not described in the report. However, with a vocabulary of just 200 words and a rule-base of 40 inference rules, the system was probably a demonstration prototype and the linguistic processing was probably tailored to handle the sample texts used.

The system developed by Joskowsicz, Ksiezyk and Grishman (1989), called *PROTEUS*, was developed to understand short narrative messages about equipment failure in Navy ships and the maintenance action performed by the crew to fix the problem. The system uses its model of the equipment to simulate the problem described in each message in order to determine whether there can possibly be a causal chain linking the events described in the message. If it fails to identify a causal chain, it can hypothesize additional facts that would enable it to complete the simulation and identify a causal chain.

The third group of studies on the extraction of causal knowledge from text dealt with expository text -- the kind of text found in textbooks. I found only two such studies, Kontos and Sidiropoulou (1991) and Kaplan and Berry-Rogghe (1991), both dealing with scientific text.

The study by Kontos and Sidiropoulou (1991) focused on causal relations between processes associated with parts of the environment -- the atmosphere, water and land. The research was a case study using a seven-page article entitled "Threats to the world's water." The approach used for identifying causal relations is similar to that used in my study, i.e. using linguistic patterns (consisting of a sequence of word and syntactic categories) to identify causal relations. However, all the information required for linguistic processing -- the grammar, the lexicon, and the patterns for identifying causal relations -- were hand-coded and were developed just to handle the seven-page article used in the study. Scaling up is obviously going to be a problem. The grammar, lexicon and patterns will not be usable in another subject area, and may not even be effective for other documents on the same subject.

The study by Kaplan and Berry-Rogghe (1991) used two sample texts on the subject of cloud formation -- one document from a children's encyclopedia and another from a handbook on meteorology. The texts were parsed by hand and converted to a propositional representation. The system used four methods for identifying causal relations:

1. using cohesive links in sentences to identify causal relations that were explicitly expressed. Cohesive links indicating cause and effect (e.g., *because*, *therefore* and *as a result*) are also used in my study, and are explained in Chapter 3 Section 3.2.
2. inferring causal relations using a knowledge of the preconditions for each event to happen. The system inferred that an event *A* caused an event *B* if *A* satisfied the preconditions for *B* to happen. This inferencing made use of hand-coded knowledge of the preconditions of events in the domain.
3. inferring causal relations using a model of the domain. The system inferred that two events were causally related if they were temporally and spatially related. A model of the domain was used to recognize that two events were spatially adjacent.
4. inferring causal relations using known constraints. This inferencing made use of mathematical equations that express quantitative relations between certain variables in the domain. The relations expressed in the equations were converted to inference rules that could be used by the system to infer cause and effect.

As the authors pointed out, substantial domain knowledge was required for the system to identify causal relations in the sample texts accurately.

My study did not make use of knowledge-based inferencing to identify causal relations, but relied entirely on linguistic clues. Knowledge-based inferencing of causal relations require a detailed knowledge of the domain. All the studies surveyed in this section dealt with very narrow domains, and most of the systems developed in the studies were demonstration prototypes working with a very small amount of text. In contrast, my study dealt with a realistic full-text database comprising about five years of *Wall Street Journal* articles. Though the *Wall Street Journal* is business oriented, it covers a very wide range of topics and the articles are non-technical. Since the purpose of this study was to develop a method for identifying causal relations that could be used by an information retrieval system dealing with a heterogeneous database, it was not possible to manually encode domain knowledge for all the subject areas covered by the database. Furthermore, I think it is important to know how effectively causal relations can be identified without the use of knowledge-based inferencing. A future study might investigate whether it is possible to use some amount of inferencing from commonsense knowledge to improve causal relation detection in a heterogeneous database.

2.3. Use of Relation Matching in Information Retrieval

2.3.1. Syntagmatic versus paradigmatic relations

Relations between terms can be divided into two types:

- syntagmatic relations
- paradigmatic relations

The distinction between syntagmatic and paradigmatic relations can be traced back to Ferdinand de

Saussure (1915/1959)[9].

A *syntagmatic relation* between two words is the relation between the words that is synthesized or expressed when a sentence containing the words is formed. The *Encyclopedia of Language and Linguistics* defines a syntagmatic relation as "the relation a linguistic unit bears to other units with which it co-occurs in a sequence or context" (Asher, 1994, v. 10, p. 5178). Lancaster (1986) characterized syntagmatic relations as *a posteriori* or transient relations.

A *paradigmatic relation* between two words is the relation between the words that is inherent in the meaning of the words. The *Encyclopedia of Language and Linguistics* defines it as the relation between linguistic units where one unit can substitute for the other according to different linguistic environments (Asher, 1994, v.10, p. 5153). Whereas syntagmatic relations are sentence dependent or document dependent, paradigmatic relations are not. Lancaster (1986) characterized paradigmatic relations as *a priori* or permanent relations. Examples of paradigmatic relations are the synonym relation, the genus-species relation (broader-narrower term), and the part-whole relation. These are relations that are typically used in a thesaurus.

Syntagmatic and paradigmatic relations are used differently in information retrieval. *Syntagmatic relations* indicate additional criteria that the retrieved document should satisfy. The retrieved document should not only contain the terms specified in the query but also express the same relations between the terms as expressed in the query. The relations are in a sense added to the search with the Boolean conjunction operator *AND*. This use of relations can be characterized as a way of *improving the precision* of the retrieval.

Paradigmatic relations are typically used for query expansion. Terms that are semantically related to each query term are added to the search using the Boolean disjunction operator *OR*. These additional *related* terms function as alternatives to the query term. Documents that do not contain a query term but contain one of the related terms instead are also retrieved in the search. This way of using relations has been characterized as *improving the recall* of the retrieval.

Is the *causal relation* a syntagmatic relation or a paradigmatic relation? The *causal relation* can be either, depending on how the relation is used, the context and the two terms involved in the relation. Harris (1987) and Gardin (1965) pointed out that when a particular relation is frequently used between two terms, the terms become permanently associated with the relation in people's memory, and the relation becomes part of the meaning of the terms. So, when a causal relation between two terms is well-established in a particular subject area and is taken for granted by the practitioners in that field, then the causal relation between the terms can be said to be a paradigmatic relation.

In this study, causal relations are used only as syntagmatic relations to increase the precision of information retrieval. The retrieval system looks for causal relations that are *expressed* in documents, and attempts to match those causal relations with the causal relations expressed in the query. Using causal relations as paradigmatic relations to expand query terms with causally related terms is left to a future study.

The following sections survey the use of relation matching for improving the precision of information retrieval. I first discuss the use of relations that are manually identified in documents, and then survey previous information retrieval research that used automatic methods for identifying syntactic and semantic relations.

[9] Saussure's used the term *associative relations* for what is now known as *paradigmatic relations*.

2.3.2. The use of manually identified relations in documents

Many conventional bibliographic databases allow the user to specify that two terms are related when searching the database subject descriptor field. These databases employ manually assigned indexing terms that are "precoordinated", i.e. some human indexer has indicated that there is a relation between the concepts in the content of the document. For most of these databases, the *type* of relation between the concepts are not specified in the indexing but implied by the context. This is the case with faceted classification schemes like Precis (Austin, 1984) and Ranganathan's Colon classification (Kishore, 1986; Ranganathan, 1965). Farradane (1967) has pointed out that the implied relations in precoordinate indexing are unambiguous only in a restricted domain. In a heterogenous database, the relation between the precoordinated terms may be ambiguous. Farradane (1967) criticized the use of *implicit* relations as being "either too vague and open to error in use, or interfere with the meaning of the concepts" (p. 298).

Two indexing systems that make *explicit* use of relations are Farradane's (1950, 1952 and 1967) relational classification system and the SYNTOL model (Gardin, 1965; Levy, 1967). Farradane used nine types of relations, and the SYNTOL project used four main types of relations that were subdivided into finer relations. In Farradane's system, the *causal relation* is subsumed under the *functional dependence* relation. In the SYNTOL system, it is subsumed under the *consecutive relation*.

It is not clear whether the use of explicit relations in indexing improves retrieval effectiveness over using keywords alone and over the use of implicit relations. The Aberystwyth Index Languages Test (Keen, 1973) found only a small improvement in retrieval precision for a minority of queries (13%) with explicit relations compared with not using relations.

2.3.3. Automatic identification of syntactic relations for information retrieval

2.3.3.1. Approaches to identifying and using syntactic relations

By *syntactic relations*, I mean the relations between terms derived from the syntactic structure of the sentence. Identification of syntactic relations in a sentence is usually done using some kind of parser to produce a syntactic parse tree. However, some systems have used simpler methods with some success. The FASIT system (Dillon and Gray, 1983) and the ALLOY system (Jones, deBessonet & Kundu, 1988) extract index phrases by looking for certain patterns of syntactic categories. The text is first tagged with syntactic category labels (i.e. part-of-speech labels), and the system then identifies sequences of syntactic categories (e.g., *adjective* followed by *noun*) that match entries in a dictionary of acceptable patterns. Every phrase that matches a pattern in the dictionary is extracted and used as an index phrase.

The Constituent Object Parser (Metzler & Haas, 1989; Metzler, Haas, Cosic & Wheeler, 1989; Metzler, Haas, Cosic & Weise, 1990) and the TINA system (Schwarz, 1990a; Schwarz, 1990b; Ruge, Schwarz, & Warner, 1991) perform syntactic processing on queries and documents to produce, not syntactic parse trees, but dependency trees that indicate which terms modify which other terms. The Constituent Object Parser project is perhaps the most ambitious retrieval system in its use of syntactic processing. The syntactic processing is customized for an information retrieval application. Ambiguities in syntactic structure that are not likely to have an impact on retrieval effectiveness are ignored. Rules were developed to handle conjunctions and ellipses. Unfortunately, a thorough evaluation of the retrieval effectiveness of the system has not been reported.

Several studies have focused on processing noun phrases only, with the assumption that noun phrases are more important than other types of phrases for information retrieval. Such is the case with the TINA system mentioned above, and the study by Smeaton and van Rijsbergen (1988).

After the syntactic relations in the query and the documents have been identified, the terms and relations in the documents have to be matched with those in the query and a score calculated to reflect the degree of match. There are two main approaches to matching terms and relations:

1. construct index phrases from the syntactic parse trees, and match the index phrases for the documents with those for the query
2. match the syntactic tree structures produced from the documents with that produced from the query statement.

The first approach of generating index phrases from the output of the linguistic processing is more commonly used. If the syntactic relations are identified by looking for sequences of syntactic categories, as in the FASIT system (Dillon and Gray, 1983), then the output already consists of phrases. The phrases have to be normalized so that different syntactic structures indicating the same relations are transformed into a canonical form. For example, the phrase *retrieval of information* might be transformed into the phrase *information retrieval*.

If the output from linguistic processing is a parse tree or some other kind of graphical representation indicating the syntactic structure (e.g., a dependency tree), then the system can apply some rules to construct a set of index phrases from the syntactic tree. In the study by Smeaton and van Rijsbergen (1988) and the study by Strzalkowski, Carballo and Marinescu (1995), pairs of terms with a dependency relation (i.e. where one term syntactically modified the other in the sentence) were extracted from the parse trees and used as content indicators. These term pairs can be considered to be pseudo-phrases.

The index phrases thus constructed can be handled simply as multi-word terms when matching documents with the query. The same procedure used for matching single-word terms can be applied to the multi-word terms. For example, an information retrieval system using the vector space model can treat the index phrases as terms in a vector.

Metzler and Haas (1989) and Sheridan and Smeaton (1992) pointed out that information is usually lost when a tree representation of a sentence structure is converted into index phrases. As an example, suppose that a document contains the noun phrase *patent information retrieval*. If two index phrases *patent information* and *information retrieval* are formed from the original noun phrase, then the document will not be retrieved if the query asks for *patent retrieval*. On the other hand, if the structure

patent -> information -> retrieval
(*patent* modifies *information* which modifies *retrieval*)

is preserved for matching. Then the system can take into account the "distance" in the relation between two terms to obtain a partial match with

patent -> retrieval

Breaking the original noun phrase structure into two index phrases also makes it impossible to distinguish between a document that contains the full phrase *the retrieval of patent information* and a document in which the phrases *patent information* and *information retrieval* appear separately but not together as one phrase, as in this example:

. . . retrieval of information about patent information offices.

Clearly, information is lost when a tree structure is broken up into smaller fragments and represented as phrases.

Retaining the tree representation of the document sentences and the query statement allows more flexibility in matching documents with the query. Several researchers have developed and used graph-matching[10] techniques for information retrieval (Liddy & Myaeng, 1994; Lopez-Lopez, 1995; Lu, 1990; Metzler, Haas, Cosic & Weise, 1990; Myaeng & Liddy, 1993; Schwarz, 1990a; Sheridan & Smeaton, 1992). However, which graph-matching procedure is best for information retrieval purposes has yet to be determined by researchers.

Although this study uses a semantic relation rather than syntactic relations, the representation and matching scheme adopted for this study are similar to the approach of generating index phrases from relations and handling them as terms in a vector. In this study, each pair of causally related terms in the query or document is treated as a term in the vector-representation of the query or document. In effect, a pseudo-phrase is formed from every pair of terms that are causally related.

2.3.3.2. Does the use of syntactic relations improve retrieval effectiveness

Research to date has found a small improvement in retrieval effectiveness when syntactic relations in documents and queries are taken into account in the retrieval process (Croft, 1986; Croft, Turtle & Lewis, 1991; Dillon & Gray, 1983; Smeaton & van Rijsbergen, 1988). The improvement over using just keywords is usually less than 10% in the *11-point recall-precision average*[11]. Strzalkowski, Carballo and Marinescu (1995) obtained an improvement of 20%, but their system included other enhancements than relation matching. Smeaton, O'Donnell and Kelledy (1995) obtained worse results from relation matching (using a tree-matching procedure) than from keyword matching.

The retrieval performance from syntactic relation matching appears to be no better than and often worse than the performance obtainable using index phrases generated using statistical methods, such as those described in Salton, Yang and Yu (1975) and Fagan (1989). Fagan (1989), who used a non-syntactic procedure for constructing index phrases based on term specificity, co-occurrence frequency and proximity, obtained retrieval improvements ranging from 2 to 23% depending on the document collection.

In comparing statistical and syntactic methods for generating index phrases for back-of-book indexing, Salton, Buckley and Smith (1990) found that the *proportion of acceptable index phrases* obtained by the best statistical and the best syntactic methods were about the same, although the syntactic method generated slightly more phrases. They concluded that the small improvement did not justify using syntactic methods as they were more complex and used more computer resources than statistical methods.

There are a few possible reasons why syntactic relation matching has not yielded better results than using word co-occurrence data. One is the difficulty of getting correct parses for most of the sentences in a database. With the development of better parsers, bigger retrieval improvements can be expected.

There are other ways in which past studies have been less than optimal in their use of syntactic relation matching. The FASIT system (Dillon & Gray, 1983) and the ALLOY system

[10] A tree is one type of graph.

[11] This retrieval effectiveness measure is explained in Chapter 5.

(Jones, deBessonet & Kundu, 1988) do not syntactically parse the sentences, but instead extract index phrases by looking for certain sequences of part-of-speech labels that have been assigned to text words. Croft, Turtle and Lewis (1991) performed syntactic processing on the queries but not on the documents. In one of their experiments, a phrase from the query was assumed to occur in the document if all the words in the phrase occurred somewhere in the document. In another experiment, the words in the query phrase were required to occur in close proximity in a document sentence. Similarly, Smeaton and van Rijsbergen (1988) performed syntactic processing on the queries but not on the documents. Each pair of syntactically related terms from the query was assumed to occur in the document if they co-occurred within the same sentence in the document. Also, Smeaton and van Rijsbergen processed only noun phrases. Smeaton, O'Donnell and Kelledy (1995) used tree-matching to determine the degree of match between a document and the query. The fact that their tree-matching approach gave worse results than a keyword matching scheme suggests that either their method of tree-matching or their method of scoring the document-query match was not optimal for information retrieval.

Fagan (1989) used a statistical method for generating index phrases. From his analysis of the errors in phrase construction, he concluded that index phrases of much better quality could be constructed if syntactic information was incorporated into the phrase construction procedure.

Croft, Turtle and Lewis (1991) observed that the effect of using syntactic relations appeared to increase with the size of the database. The use of syntactic relations allows finer distinctions to be made between documents and this is likely to become more apparent the larger the document collection.

2.3.4. The use of semantic relations

By *semantic relations*, I mean the logical or conceptual relations expressed in the text. A semantic relation is partly but not entirely determined by the syntactic structure of the sentence. A particular semantic relation can be expressed using various syntactic structures. Systems that use automatic methods to identify semantic relations in text usually do it as part of the process of extracting information to store in a semantic representation (or knowledge representation scheme). Natural language text that has been converted to a semantic representation is easier to manipulate by a computer. Information is retrieved from this semantic store by comparing the information in the store with the semantic representation of the user's query. Such a system is called a *conceptual information retrieval system*.

To extract information from text requires extensive domain knowledge to support the syntactic and semantic processing. Since the knowledge base has to be constructed manually, such systems are usually limited to a narrow domain. In a specialized technical domain, sentence structures may show less variety than in a non-technical or heterogeneous document collection. Moreover, the words used in a specialized technical domain may be limited to a relatively small technical vocabulary. Syntactic processing can thus focus on the common syntactic structures, and the knowledge base construction and semantic processing can focus on the terms and concepts that are important in that domain.

Four examples of *conceptual information retrieval systems* are the RIME system (Berrut, 1990), the patent-claim retrieval system described by Nishida and Takamatsu (1982), the SCISOR system (Rau, 1987; Rau, Jacobs & Zernik, 1989) and the FERRET system (Mauldin, 1991). The RIME system is used for retrieving X-ray pictures each associated with a medical report describing in natural language the content and medical interpretation of the picture. The system automatically converts the medical reports to binary tree structures that represent medical concepts and relations between them. The patent-claim retrieval system described by Nishida and Takamatsu (1982) extracts information from patent-claim sentences in patent documents and stores the information in a relational database. The SCISOR system extracts information from short newspaper stories in the domain of corporate takeovers, and stores the information in the KODIAK knowledge representation language -- a hybrid frame and semantic net-based representation formalism. The FERRET system has been used to convert astronomy texts to a frame representation.

The procedures used by these systems to identify semantic relations between terms are too complex to describe here. However, all these systems use extensive domain-specific knowledge about what relations tend to occur in specific situations and with particular types of concepts. One common way of storing domain knowledge is in the form of *case frames* that specify the participant roles in an event, what types of entities can fill those roles and what syntactic function each participant will have in the sentence (Fillmore, 1968; Somers, 1987). During semantic processing, case frames are usually triggered by verbs that indicate a particular event, but they may be triggered by other words as well. The SCISOR, FERRET and the patent-claim system mentioned above all use case frames. In addition, these systems also use higher-level knowledge represented as scripts. A *script* specifies a typical sequence of events, the relation between events and the relation between the participant roles of the various events in the script.

It is not clear how effective these systems are. The system evaluations have not been carried out in a way that allows comparison with the best keyword matching methods.

Lu (1990) investigated the use of *case relation* matching for information retrieval using a small test database of abstracts (mostly from the ERIC database). The retrieval results that he obtained from case relation matching were worse than from vector-based keyword matching (though not significantly so). He used a tree-matching technique for matching case relations. It may not be fair to compare a relation matching scheme that uses tree matching with a keyword matching scheme that uses vector matching. Vector-based keyword matching has been studied for decades and the best methods are known. Research on tree-matching methods for information retrieval has barely begun. I think a fairer comparison in Lu's (1990) study would have been between tree matching with relation and tree matching assuming there were no relation matches.

Gay and Croft (1990) studied one difficult aspect of semantic processing -- the identification of semantic relations between the members of compound nouns. *Compound nouns* (also called *compound nominals* or *nominal compounds*) is a sequence of two or more nouns forming a unit that itself acts as a noun. Some examples are *college junior*, *junior college*, and *information retrieval*. The authors found that commonsense knowledge is essential for interpreting compound nouns. The knowledge base they used included case frames described earlier, information about which entities tend to be associated with each event (e.g., *food* tends to be associated with the *eat* event), information about what participant roles an entity tends to be associated with (e.g., *screwdriver* tends to have a role of *instrument*), and a hierarchical classification of entities. In a small experiment using the CACM document collection (Fox, 1983), the authors found that although their knowledge intensive procedure correctly interpreted compound nouns about 76% of the time, it was not likely to yield a substantial improvement in retrieval effectiveness.

Identifying and coding the necessary domain knowledge for semantic processing is labor-intensive and time-consuming. Moreover, much of the knowledge is not portable to another domain. It is thus important to investigate whether non-domain specific procedures for extracting semantic relations can yield a substantial improvement in retrieval effectiveness.

The DR-LINK project (Liddy & Myaeng, 1993; Liddy & Myaeng, 1994; Myaeng & Liddy, 1993; Myaeng, Khoo & Li, 1994) was perhaps the first large-scale project to investigate general methods for extracting semantic relations for information retrieval. Non-domain specific resources like the machine readable versions of *Longman Dictionary of Contemporary English* (2nd ed.) and *Roget's International Thesaurus* (3rd ed.) were used. Case frames were constructed semi-manually for all verb entries and senses in the *Longman Dictionary*. Though case frame construction involved substantial manual effort, the case frames were not limited to a particular domain. Besides case frames, lexical patterns were used to identify additional semantic relations. However, preliminary experiments found few relation matches between queries and documents.

From my experience with the DR-LINK project and the literature survey given in this chapter, I surmise that better retrieval results can be obtained by focusing research efforts on particular types of semantic relations (e.g., *agent* or *causal* relation) rather than on particular syntactic units (e.g., noun phrases or compound nouns) or particular methods (e.g., using case frames). A semantic relation can be expressed in many syntactic forms. The following examples show several ways in which the *agent* or *actor* relation between *John* and the action of *driving* can be expressed in the English language:

John (agent) drove the van.
John (agent) was spotted driving the red van.
John (agent) was the driver of the van.
The driver, John (agent), was given a speeding ticket.

By focusing on one type of semantic relation at a time, we can employ a combination of methods to process all the syntactic constructions necessary to identify most instances of the relation in text. We can also investigate the effects of each type of semantic relation on retrieval effectiveness. This study is an in-depth study of one such semantic relation -- the causal relation.

CHAPTER 3
EXPRESSION OF CAUSAL RELATION IN TEXT

3.1. Introduction

This chapter surveys the various ways in which the causal relation is expressed in written English. There are probably other ways in which people express cause and effect in spoken discourse but little has been done to investigate this. In any case, this study dealt only with causal relations in published text.

Many of the causal relations in text are implicit and are inferred by the reader using general knowledge. Consider the following two examples:

(1a) John had a car accident. He was taken to the hospital in an ambulance.
(1b) John had a car accident, and was taken to the hospital in an ambulance.

Although the causal relation is not explicitly indicated in either of the examples, the reader has no difficulty inferring that the car accident caused John to be taken to the hospital. Furthermore, from the fact that an ambulance was involved, one can infer that John suffered a serious injury as a result of the accident. This chapter does not discuss how cause and effect is inferred during text comprehension, but only how cause and effect is explicitly indicated in written English.

I first describe Altenberg's (1984) *typology of causal linkage* which covers linking words used to indicate a causal relation between clauses or phrases, and to indicate which clause or phrase is the cause and which the effect. The nature of the linking words is described. I then discuss ways of indicating cause and effect not covered by Altenberg's typology, such as the use of causal verbs, adverbs and adjectives, the use of conditionals and the implicit attribution of cause by verbs.

3.2. Altenberg's Typology of Causal Linkage

In the examples given below, "C" represents *cause* and "E" represents *effect*.

TYPE OF LINK	CAUSE-EFFECT ORDER	EFFECT-CAUSE ORDER
A. ADVERBIAL		
1. Anaphoric (conjunct)		
a. implicit cohesion	C, so E	-
b. pronominal cohesion or determiner	C. For this reason (NP) E	-
c. pronominal and lexical cohesion	C, owing to NP E	-
2. Cataphoric		
a. appositive	C with the result that E	E, for the reason that C
b. prospective	C with these consequences: E	E for X reasons: C
B. PREPOSITIONAL		
1. Adverbial member	Because of C, E	E, because of C
2. Postmodifying member	-	E(NP) due to C
C. SUBORDINATION		
1. Subordinator	C, so that E	E, because C
2. Non-finite *ing*-clause	Being C, E	E, being C
3. Correlative comparative	C so Adj . . . that E	E . . . so Adj C
D. CLAUSE-INTEGRATION		
1. Rhematic link (link as complement)		
a. clause internal	-	E is due to C
b. anaphoric pronominal subject	C. This is the reason why E	E. This is the consequence of C
c. anaphoric nominal subject	C. NP is the cause of E	E. NP is due to C
d. non-anaphoric subject	C. NP came to the conclusion that that was why E	-
e. cleft construction	-	E. It is because C that (E)
f. retrospective reinforcement	-	E . . . C. That's why (E)
g. mediating	C. NP draws his conclusions: E	E. There are reasons for this:C
h. prospective (thematic first member)	C(NP) has two Adj effects: E	-
i. prospective (rhematic first member)	-	There are X grounds for E(NP). C
2. Thematic link (link as subject)		
a. anaphoric	C. The result is E	E. The reason is C
b. mediating	C. The result is Adj: E	E. The reason is Adj: C

Figure 1. Altenberg's (1984) typology of causal linkage.

Altenberg's (1984) typology is summarized in 1. Altenberg classified causal links into four main types:

1. the adverbial link, e.g. *so, hence, therefore*
2. the prepositional link, e.g. *because of, on account of*
3. subordination, e.g. *because, as, since*
4. the clause-integrated link, e.g. *that's why, the result was.*

Altenberg's typology is restricted to explicit (i.e. overtly realized) links, and does not include causal relations lexicalized as verbs, e.g. *cause, make, produce, result in/from,* and *follow (from)*.

3.2.1. Adverbial linkage

An *adverbial link* is an adverbial which provides a cohesive link between two clauses. The link may be between two sentences. *Adverbial linkage* is subdivided into:

1. the *anaphoric adverbial*, which has an anaphoric reference to the preceding clause, e.g.
   (2) There was a lot of snow on the ground. *For this reason* the car failed to brake in time.

2. the *cataphoric adverbial*, which has a cataphoric reference to the following clause, e.g.
   (3) There was a lot of snow on the ground *with the result that* the car failed to brake in time.

With *anaphoric adverbials*, the cohesive reference can be:

a. *implicit*, i.e. the anaphoric reference is not overtly expressed (e.g., *so, consequently, as a result*).
b. *pronominal or a determiner*, i.e. the reference is explicitly marked by an anaphoric pronoun or determiner (e.g. *as a result of that, because of this, for this reason*).
c. *pronominal and lexical*, where the reference is further specified by a noun phrase which summarizes the content of the preceding clause, e.g.
   (4) It was snowing heavily, and because of the snow the car didn't brake in time.

With *cataphoric adverbials*, the second clause is either:

a. attached to the linking words as an *appositive clause*, e.g.
   (5) There was a lot of snow on the road *with the result that* the car failed to brake in time.
b. presented in a new independent clause (*prospective linking*):
   (6) There was an unexpected snow storm over the holiday weekend, with the following consequences: . . .

3.2.2. Prepositional linkage

Whereas an *adverbial linkage* (described in the previous section) links two clauses, a *prepositional linkage* connects cause and effect in the same clause. The prepositional phrase formed usually has an adverbial function, i.e. the preposition links a noun phrase to the clause, for example:

(7) The car failed to brake in time *because of* the slippery road.

In the above example, the words "because of" function as a phrasal preposition.

Occasionally, the prepositional phrase modifies a noun phrase, as in this example:

(8) The car crash, *due to* slippery road conditions, could have been avoided had the road been cleared of snow.

3.2.3. Subordination

In linkage by subordination, the link can be:

1. a subordinator (e.g. *because, as, since, for, so*)
2. a structural link marked by a non-finite *ing*-clause, as in the following example:
   (9)*Being* wet, the road was slippery.
3. a correlative comparative construction (e.g. *so . . . that/as to, too . . . to*)
   (10) There was *so* much snow on the road *that* the car couldn't brake in time.

3.2.4. Clause integrated linkage

Clause-integrated links are so named because they form part of the *subject* or the *predicative complement* of a clause. Unlike the previous three types of links that have peripheral syntactic function in their clauses, clause-integrated links operate as central clause constituents.

When the linking words are the *subject* of a clause, Altenberg called it a *thematic link*. The following sentence has a *thematic link*:

(11) The car didn't brake in time. *The reason* was that there was a lot of snow on the road.

The linking words "the reason" function as the subject of the sentence.

When the linking words form part of the *predicative complement* of a clause, it is called a *rhematic link*:

(12) The car accident was *due to* the slippery road.
(13) There was a lot of snow on the road. This was *(the reason) why* the car didn't brake in time.

In each of the above examples, the linking words ("due to" in example (13), and "the reason why" in example (14)) form the first part of the phrase that functions as the complement of the copular verb.

Altenberg further divided the *thematic link* into two subtypes, and the *rhematic link* into nine subtypes. These are listed with examples in 1.

3.3. Causal Verbs

Causal verbs (also referred to as *causative verbs* or *lexical causatives*) are verbs the meanings of which include a causal element. Examples include the transitive form of *break* and *kill*. The transitive *break* can be paraphrased as *to cause to break*, and the transitive *kill* can be paraphrased as *to cause to die*.

Thompson (1987) divided causal verbs into three groups, examples of which are given here:

Causal verbs	Phrasal equivalents
Group 1	
x breaks y	x causes y to break
x moves y	x causes y to move
x melts y	x causes y to melt
Group 2	
x kills y	x causes y to die
x convinces y	x causes y to believe
x raises y	x causes y to rise
Group 3	
x butters y	x causes y to be buttered
x ties y	x causes y to be tied
x hides y	x causes y to be hidden

In *group 1*, each transitive causal verb also has an intransitive usage.[12] For example, "x *breaks* y" is paraphrased as "x causes y to *break*." This is not true of verbs in the *group 2*. The transitive *kill* is paraphrased using the intransitive *die*, a different word. For verbs in *group 3*, there is no intransitive verb that can be used in the paraphrase. The past participle form of the causal verb is used instead. "X butters y" is paraphrased as "x causes y to be *buttered*."

Several writers have pointed out that the *causal verb* is not synonymous with the *causal paraphrase*. "X kills y" is not exactly the same as "x causes y to die." Whereas "x kills y" entails "x causes y to die," "x causes y to die" does not entail "x kills y." This can be seen by connecting the two forms by *but* with one sentence in the affirmative and the other in the negative as suggested by Shibatani (1976):

(14a) * John killed Bill, but he didn't cause him to die.[13]
(14b) John caused Bill to die, but he didn't kill him.

As an example, if *x* thoughtlessly dropped a banana peel on the floor and *y* slipped on it and died from his injuries, one would say that *x* caused *y*'s death but not that *x* killed *y*.

[12] The phenomenon where an intransitive verb also has a transitive usage with the meaning "cause to *verb-intransitive*" has been referred to as causative alternation. A few writers have sought to explain why certain verbs have this property. Levin & Hovav (1994) argued that the intransitive verbs that also have a transitive usage with a causal meaning are those verbs that denote events which are *externally caused*. "Verbs which are externally caused inherently imply the existence of an external cause with immediate control over bringing about the eventuality denoted by the verb." (Levin & Hovav, 1994, p. 50) Not all transitive causal verbs have an intransitive usage with the object of the transitive verb occupying the subject position of the intransitive verb. For example, the verbs *assassinate* and *build* cannot be used in this way. Levin & Hovav (1994) argued that these are verbs that denote eventualities that can only be caused by an animate volitional agent. Transitive causal verbs that also have an intransitive usage (e.g. *break*) are those that need not be caused by a volitional agent, and can be caused by a natural force, an instrument or some circumstance.

[13] An asterisk before an example sentence indicates that the sentence is ungrammatical.

How is the meaning of the *causal verb* different from that of its *causal paraphrase*? Gonsalves (1986) suggested that the difference is in the degree of directness with which the agent is involved in the causation. The *causal verb* indicates that the agent (i.e. the subject of the verb) "participates crucially in the causation by his acts or act." Thompson (1987) argued to the contrary that each verb allows particular ways of performing the action meant by the verb, and the acceptable ways are more direct for some verbs than for others. To *walk* a dog, one pretty much has to walk with the dog (in addition to making the dog walk). On the other hand, there are many more ways to *convince* a person of something, some of which can be indirect. Thompson pointed out that even if one could *directly* cause someone to believe something by giving him a pill, it would still not be considered as *convincing* the person. *Convincing* someone has to involve engaging the person's mental processes. Said Thompson (1987):

> Each causal verb trails a very fuzzy bounded array of appropriate causal routes, loosely fixed by our interests . . . and by our customary ways of doing things. (p. 107)

Kovalyova (1979) pointed out that *kill* can be synonymous with *cause to die* in certain contexts. She argued that the meaning of *kill* has both a physical action component and a causative component. In some contexts the physical action meaning is dominant whereas in other contexts the causative meaning is dominant.

For the purpose of this study, it is sufficient that "x kills y" implies that "x causes y to die" even if the two constructions do not always have exactly the same meaning. For each occurrence of a causal verb (e.g., *kill*) in the text, this study inferred that there was a causal relation between the subject of the verb and some event (e.g., *dying*) involving the object of the verb.

Another important issue is how to distinguish causal verbs from other transitive verbs that are not causal. It may be argued that all transitive action verbs are causal since an action verb such as *hit* can be paraphrased as *to cause to be hit*. Indeed, Lyons (1977, p. 490) said that there is a natural tendency to identify causality with agency and that causativity involves both causality and agency. Wojcik (1973, p. 21-22) said that "all agentive verbs involve the semantic prime CAUSE at some level." Agents may be said to "cause" themselves to do things. The sentence "John intentionally broke the window" seems to entail "John caused himself to break the window." Wojcik considered all action verbs to be causal verbs in this sense.

However, most writers on the topic of causal verbs do not equate action verbs with causal verbs. Thompson (1987) argued that verbs like *hit*, *kick*, *slap*, and *bite* are not causal. She said that whereas causal verbs accept events and states of affairs as subjects, verbs like *hit* do not. Consider the following sentences:

(15a) Oswald killed Kennedy by shooting at him.
(15b) Oswald's accurate shooting killed Kennedy.

(16a) John broke the vase by shooting at it.
(16b) John's accurate shooting broke the vase.

(17a) Tom hit the can by shooting at it.
(17b) * Tom's accurate shooting hit the can.

Examples (15b) and (16b) show that the subject of the verbs *kill* and *break* can be an event. Examples (17a) and (17b) show that the subject of the verb *hit* can only be a person or an object. Thompson said that this is because for action verbs like *hit* the subject is, in a sense, not separable from the result of the action.

Vendler (1984) and Szeto (1988) drew the distinction between *actions* and *events*.

Actions are done or performed by an agent, whereas events are caused. *Agent* is related to *action*, whereas *cause* is related to *event*. Hence, *agent* plays no direct role in the causation of an *event*. It is the *action performed by the agent* that can cause an event. Szeto (1988) used the distinction between events and actions to distinguish between *event verbs* and *action verbs*. *Event verbs* (e.g., *break*, *open*, and *melt*) do not require an agent to be specified and can thus have an intransitive usage in which the subject of the verb is not the agent but is the patient of the verb. On the other hand, *action verbs* (e.g., *hit*, *cut*, and *chew*) normally require an agent to be specified in the clause.[14]

Szeto also claimed that *event verbs*, like *break*, are vague about the precise nature of the action involved but specific about the results. On the other hand, *action verbs*, like *hit*, are precise about the action but "completely noncommittal as to the nature of the result of the action." If the result of an *action verb* is to be explicitly pointed out, it is usually done by adding a "resultative" phrase as in:

(18) John *cut* the meat in half/into slices.

Szeto equated his *event verbs* with *causal verbs*[15]. He classified causal verbs into the following types:

- Verbs of breaking and destroying, e.g. break, destroy
- Verbs of moving and stopping: move, stop
- Verbs of opening and closing: open, close
- Verbs of changing: change
- melt, grow, freeze, boil, burn, cool, heat[16]
- Verbs of cooking: bake, cook, broil, grill, stew
- De-adjectival verbs: darken, blacken, sharpen, widen
- Verbs of locomotion: walk, gallop, march, swim, jump

Comparing Szeto's characterization of causal verbs with Thompson's three groups of causal verbs described earlier, it can be seen that only group 1 of Thompson's causal verbs satisfies Szeto's criterion of *event verbs*. Only group 1 verbs have an intransitive usage.

To recapitulate, the following criteria have been proposed for distinguishing causal verbs from other transitive action verbs:

1. *Causal verbs accept events and states of affairs as subjects, whereas other action verbs accept only agents as subjects* (Thompson, 1987). Thompson said that this is because for action verbs, the subject is, in a sense, not separable from the result of the action. Although this criterion is promising, it was not adopted for this study because the criterion had not been used on a large scale to identify causal verbs, and it was not known whether this test would yield intuitively acceptable results. One difficulty with this test is that to rule out a verb as causal, one has to be certain that there is no event or state that can be used as the subject of the verb. Also, the way the event or state is expressed may determine whether it is acceptable as the subject of the verb. It will take a considerable amount of time and effort to apply the

[14] Exceptions are constructions such as:
The meat *cuts* easily.

[15] Actually, Szeto distinguished between two kinds of causal verbs, *event verbs* and *process verbs*. He considered the verbs of locomotion like *march*, *walk* and *jump* to be *process verbs*.

[16] Szeto didn't provide a label for this group of verbs. Also, it is not clear how this group defers from *verbs of changing*.

test to all the verb entries in a dictionary.

2. *Causal verbs are transitive verbs which also have an intransitive usage where the subject of the verb has the patient role* (Szeto, 1988). This criterion is too narrow because it excludes verbs like *kill*, which is usually considered to be a causal verb.
3. *Causal verbs specify the result of the action, whereas other action verbs specify the action but not the result of the action* (Szeto, 1988).

For the purpose of this study, I took the third criterion as a working definition of a causal verb: a causal verb is a transitive verb that specifies the result of an action, event or state, or the influence of some object. Causal verbs include some action verbs like *kill*, as well as some transitive verbs like *amaze* which are not action verbs but nevertheless specify the impact of some object or event.

Wojcik (1973) pointed out that there is a special class of verbs that are primarily causal in meaning. He divided them into:

- coercive causatives, e.g. force, coerce, compel
- neutral causatives, e.g. cause, result in, result from, lead to
- permissive causatives, e.g. allow, let, permit

I included all these as causal verbs.

I also made use of the rule (adapted from Thompson, 1987) that the subject of a causal verb must be separable from the result. This is to exclude words like *mar*, *surround* and *marry*, for which the subject of the verb is an integral part of the effect specified by the verb, as in the following examples from *Longman Dictionary of Contemporary English* (2nd ed.):

(19) The new power station *mars* the beauty of the countryside.
(20) A high wall *surrounds* the prison amp.
(21) Will you *marry* me?

To obtain a comprehensive list of causal verbs, the above criteria were applied to the first two senses of all the verb entries in the *Longman Dictionary of Contemporary English* (2nd ed.). In doing so, I encountered two problems.

The first was that transitive verbs specify the result of an action to a greater or lesser extent. The degree to which verbs specify the result of an action is a continuum. It is not possible to classify verbs cleanly into verbs that specify the result of an action and verbs that do not. Even the verb *break*, which is used by many writers as an example of a causal verb, does not completely specify its result. In the sentence

(22) John broke the vase.

it is not specified whether the vase was broken in two or into a thousand pieces. In classifying verbs into causal and non-causal, I accepted a verb as causal even when the result was only vaguely specified.

The second problem I encountered was in deciding what counted as a "result." Does it refer only to changes in physical state, or does it include mental states as well. What about changes in the location of an object? Whether a verb is seen as specifying a result probably depends on the domain of application and the perspective of the reader. Take the verb *hit* which is usually considered to be non-causal. In the sentence

(23) John hit Ted with a bat.

we don't know whether Ted suffered a bruise or a broken leg as a result of John's hitting. However, we do know that there is forceful contact between the bat and Ted. So, one can argue that the result of John's action is contact between the bat and some part of Ted's body. For the purpose of information retrieval, it is probably not so important what is accepted as "result" than that it is done consistently. To help make consistent decisions on whether a verb was causal or not, I identified 47 types of results. These are listed in Appendix 3, together with a list of causal verbs for each type of result. A verb was accepted as causal if it specified one of the listed types of results. Verbs that do not belong to one of the 47 types but are nevertheless clearly causal are listed in the "miscellaneous" category A43.

As mentioned earlier, in constructing the list of causal verbs in Appendix 3, the first two senses of each verb entry in the *Longman Dictionary* were examined.[17] Other senses were also considered if they could be distinguished syntactically from the first two senses. For example, if the first two senses of a verb are usually not used with a prepositional phrase, whereas the fourth sense is usually used with a particular preposition, then the fourth sense of the verb was also considered for entry in the list of causal verbs. The fourth sense of the verb is distinguishable from the first two senses by the prepositional phrase. Also considered for entry in the list of causal verbs was the first sense of each phrasal verb entry in the *Longman Dictionary*.[18] Some verbs don't have a separate verb entry in *Longman Dictionary* but are listed within the entry for a noun or adjective, i.e. they are indicated as being derived from the noun or adjective. Verbs that don't have a separate verb entry in *Longman Dictionary* were not scanned for inclusion in the list of causal verbs. However, for verbs in the Wall Street Journal document collection that don't appear as a verb entry in the *Longman Dictionary*, *Webster's Third New International Dictionary of the English Language* was consulted to determine if they were causal verbs.[19] There are a total of 2082 unique verbs[20] in the list of causal verbs in Appendix 3.

3.4. Resultative Constructions

A resultative construction is a sentence in which the object of a verb is followed by a phrase describing the state of the object as a result of the action denoted by the verb. The following examples are from Simpson (1983):

(24a) I painted the car *yellow*.
(24b) I painted the car *a pale shade of yellow*.
(24c) I cooked the meat *to a cinder*.
(24d) The boxer knocked John *out*.

[17] In Longman Dictionary, the most common senses are listed first in each entry.

[18] A phrasal verb is a sequence of two or more words having the function of a verb.

[19] Verbs in the Wall Street Journal Collection were identified using the POST part-of-speech tagger obtained from BBN Systems and Technologies.

[20] Some verbs appear more than once in the list because they have more than one sense with a causal meaning.

In example (24a), the adjective *yellow* describes the color of the car as the result of the action of painting the car. This phrase that describes the result of the action denoted by the verb is called a *resultative attribute* or *resultative phrase*. In each of the four examples above, the phrase in italics is the resultative phrase. The examples show that a resultative phrase can be:

- an adjective, as in example (24a)
- a noun phrase, as in (24b)
- a prepositional phrase, as in (24c)
- a particle, as in (24d).

Simpson (1983) described various kinds of resultative constructions. I briefly highlight the features of some of them. Simpson showed that some verbs that are normally intransitive can take an object if followed by a resultative phrase:

(25a) He shouted.
(25b) * He shouted himself
(25c) He shouted himself *hoarse*.

(26a) I cried.
(26b) * I cried myself.
(26c) I cried myself *to sleep*.

(27a) * I cried my eyes.
(27b) I cried my eyes *blind*.

The verbs *shout* and *cry* are intransitive verbs, as shown in examples (25a-b), (26a-b) and (27a). Examples (25c), (26c) and (27b) show that these verbs can be followed by an object and a resultative phrase. The objects in these three sentences have been called *fake objects* (Goldberg, 1991). In examples (25c) and (26c), the object is a reflexive pronoun, whereas in example (27b) it is not. Simpson (1983) noted that instances involving reflexive objects are more common. She referred to these reflexive objects as *fake reflexives*.

There are some transitive verbs that will take as object a noun phrase that they don't normally accept as object, if the noun phrase is followed by an appropriate resultative phrase. Consider the following examples:

(28a) John drank the beer.
(28b) * John drank himself.
(28c) John drank himself *into the grave*.
(28d) * John drank me.
(28e) John drank me *under the table*.
(28f) * John drank his family.
(28g) John drank his family *out of house and home*.

In examples (28c), (28e) and (28g), the object of the verb *drink* is not the "patient" of the verb, i.e. it does not denote the thing that John drank.

Resultative constructions are not limited to sentences where the verb is followed by an object. The following two sentences are also considered by linguists to be resultatives:

(29) The ice cream froze *solid*.
(30) The ice cream was frozen *solid*.

These sentences specify the result but not the cause, i.e. they don't say what caused the ice cream to become solid. Such resultatives are not relevant to this study since this study is concerned with sentences that specify both cause and effect.

An important question is whether all verbs will take an appropriate resultative phrase. In other words, for each verb, is there some resultative phrase (possibly one that nobody has thought of yet) that the verb will accept? This question is difficult to answer. The fact that one can't think of an "acceptable" resultative phrase for a verb does not mean there isn't one. And if it is true that some verbs can take a resultative phrase and other verbs can't, then how can these two classes of verbs be differentiated? Is there a systematic explanation for these two classes of verbs? These questions have not been satisfactorily answered by linguists.

Simpson (1983) suggested that there are at least two semantic constraints on resultative attributes in the English language:

1. the verb that accepts a resultative phrase must denote an action that "necessarily" affects the object of the verb
2. the verb that accepts a resultative phrase cannot denote a change of location.

Classes of verbs that satisfy the constraints include verbs of contact (e.g., *hit*) and verbs of change of state (e.g., *freeze*). According to Simpson, classes of verbs that do not satisfy the constraints and, thus, cannot take a resultative phrase include verbs of perception (e.g., *see*) and verbs of change of location (e.g., *fall* and *send*).

Hoekstra (1988) pointed out that the meaning of "affect" in Simpson's first constraint is ambiguous. The following sentences are resultatives, yet the action denoted by the verb does not act on the object in a direct way:

(31) He laughed himself *sick*.
(32) She laughed him *out of his patience*.
(33) The clock ticked *the baby awake*.

The verbs *laugh* and *tick* normally don't even take an object. Hoekstra (1988) claimed that all verbs that can take a resultative phrase are non-stative verbs (i.e. verbs denoting an activity or process), but that not all non-stative verbs can take a resultative phrase:

> In principle, each non-stative verb may appear with a result denoting SC [small clause], but in fact the distribution appears to be more restricted, showing that language does not fully exploit its resources. (p. 138)

It is not known how the non-stative verbs that may take a resultative phrase can be differentiated from the non-stative verbs that cannot take a resultative phrase. Clearly, more work needs to be done to arrive at a satisfactory explanation of why some verbs can take a resultative phrase and other verbs can't.

Even when a verb can take a resultative phrase, it usually will not take just any kind of object and resultative phrase, even if the resultative phrase is meaningful:

(34a) He shouted himself hoarse.
(34b) * He shouted me deaf. (i.e. the shouting caused me to become deaf)

(35a) I cooked the meat to a cinder.
(35b) * I cooked the meat burnt. (i.e. the meat was burnt)

(36a) The clock ticked the baby awake.
(36b) * The clock ticked me crazy. (i.e. the ticking drove me crazy)

(37a) I cried my eyes blind.
(37b) * The baby cried her mother crazy. (i.e. the crying drove the mother crazy)

Simpson (1983) said that "verbs in general are relatively restricted as to what resultative attribute, and what object they may appear with. . . . it is likely that most such verbs will have to list as part of their lexical entry, what the FORM of the resultative and the object will be." (p.151) To my knowledge, no one has attempted to identify general principles that determine the form and the meaning of the resultative phrase that a verb can take.

The syntactic structure of resultative sentences is still a matter for debate. A recent treatment is found in Carrier and Randall (1992). I shall not review the issues relating to the syntactic structure of such sentences. In this study, my concern is with how resultative sentences can be identified in text, and how such sentences can be distinguished from other sentences that are not resultatives but have the same syntactic pattern (i.e. have the same string of syntactic categories).

I mentioned earlier that a resultative phrase can be an adjective, a noun phrase, a prepositional phrase or a particle. This gives us four types of verb phrase structures: V-NP-Adj, V-NP-NP, V-NP-PP and V-NP-Particle. As the examples below illustrate, non-resultative sentences can have these syntactic patterns too:

not relevant !

V-NP-Adj (examples from Rapoport (1990))
Roni ate the meat raw. (depictive)
The children found Puddlegum interesting. (small clause)
At midnight, Caspian saw the children upset. (perception verb)[21]

V-NP-NP
I painted him a picture of his home.

V-NP-PP
John drank whiskey under the table.
John drank whiskey out of the bottle.

V-NP-Particle
I found him out.

So, resultative constructions cannot be identified reliably using syntactic patterns alone. How resultatives can be identified in text and distinguished from non-resultatives by a computer program has not been studied.

In this study, I make use of the syntactic pattern *V-NP-Adj* to identify resultative sentences in which the resultative phrase is an adjective. Simpson (1983) said that this is the most common kind of resultative. As I mentioned earlier, non-resultative sentences can have this syntactic pattern too. In the next chapter, I shall evaluate how much error is incurred by using this pattern to identify resultatives.

Some verb entries in *Longman Dictionary* do indicate what kind of resultative phrase can be used with the verb. For example, the entry for the verb *drink* has the grammar code "T+obj+adv/prep" indicating that the *T*ransitive verb can be followed by an *obj*ect (i.e. noun phrase)

[21] The descriptive labels given in parentheses are from Rapoport (1990).

and an *adv*erb or *prep*ositional phrase. The definition given

> *to bring to a stated condition by drinking alcohol*

suggests that the construction is a resultative. It is possible then to use the pattern

> drink <noun phrase> <prepositional phrase>

to identify such resultative sentences as (28c), (28e) and (28g), reproduced below:

(28c) John drank himself *into the grave*.
(28e) John drank me *under the table*.
(28g) John drank his family *out of house and home*.

Of the 15 examples of resultative sentences in this section, nine of them are specified in the appropriate verb entries in *Longman Dictionary*.

In this study, such information in *Longman Dictionary* is not systematically used to identify resultative sentences. Resultative sentences involving causal verbs are handled by extracting as the *effect* not just the noun phrase following the verb but also the prepositional phrases that follow the noun phrase. Resultative constructions involving non-causal verbs are not handled in this study with the following two exceptions:

- when the resultative phrase is an adjective. As mentioned earlier, the pattern *V-NP-Adj* is used to identify resultative sentences where the resultative phrase is an adjective. Simpson (1983) said that adjectives are the category most commonly used as resultatives.
- when the resultative phrase is a particle. When verb entries in *Longman Dictionary* were scanned for causal verbs, I treated each particle listed in a dictionary entry as part of a phrasal verb. So, particles that are often used as resultative attributes with particular verbs are included in my list of causal verbs as integral parts of phrasal causal verbs.

The systematic use of dictionary information to identify resultatives involving *non-causal* verbs is left to a future study.[22]

3.5. Conditionals

"If ... then ..." conditionals assert that the occurrence of an event is contingent upon the occurrence of another event. Since the contingency of one event on another suggests a causal relation between the two events, *if-then* constructions often indicate that the antecedent (i.e. the *if* part) causes the consequent (the *then* part).

It has been found that people sometimes interpret an *if-then* construction as a conditional and sometimes as a biconditional (i.e. "if and only if"), depending on the context. A conditional specifies that the antecedent is *sufficient* for the consequent to happen. A biconditional (i.e. "if and only if") specifies that the antecedent is both *necessary and sufficient* for the consequent to happen.

Whether *if-then* is interpreted as a conditional or biconditional depends on the background information available to the subject (Cummins, Lubart, Alksnis, & Rist, 1991; Hilton, Jaspars & Clarke, 1990; Rumelhart, 1979). If the subject can think of other antecedents that can lead

[22] When I was scanning verb entries in Longman Dictionary for causal verbs, I wasn't aware of the range of resultative constructions and wasn't looking out for information about them in the dictionary.

to the consequent, then the subject will interpret *if-then* as a conditional, otherwise the subject will interpret it as a biconditional. Here are two examples from Rumelhart (1979):

(38) If you mow the lawn then I will give you $5. (biconditional)
(39) If you are a U.S. senator then you are over 35 years old. (conditional)

We tend to interpret the first statement as expressing a biconditional relation ("If and only if") whereas it seems more natural to interpret the second as asserting a simple conditional relation ("If, but not only if . . .").

Another factor that influences the interpretation of *if-then* constructions is the extremity or rarity of the consequent event. The more extreme or unusual the consequent, the more likely it is that the antecedent is judged by subjects to be necessary but not sufficient (Hilton, Jaspars & Clarke, 1990). Kun and Weiner (1973) and Cunningham and Kelley (1975) found that extreme or unusual events such as passing a difficult exam or being aggressive to a social superior seem to require explanation in terms of multiple necessary conditions (e.g. working hard and being clever, being drunk and provoked). On the other hand, non-extreme and frequent events (e.g. passing an easy exam) could have been produced by many sufficient causes on their own (e.g. working hard or being clever).

Cheng and Holyoak (1985) found also that subjects were likely to interpret permissions as expressing necessary but not sufficient condition for an action. Thus if a customer is over 18 then he/she may (i.e. is allowed to) have a drink, but will not necessarily have one.

If-then constructions do not always indicate a causal relation. In the following example there is no causal relation between the "if" and the "then" part of the sentence:

(40) If you see a lightning, you will soon hear a thunder.

Though hearing a thunder is contingent on seeing a lightning, one does not cause the other. Seeing a lightning and hearing a thunder are caused by the same atmospheric event.

3.6. Adverbs and Adjectives of Causation

There are some adverbs and adjectives which have a causal element in their meanings (Cresswell, 1981). One example is the adverb *fatally*:

(41) Brutus *fatally* wounded Caesar.

This can be paraphrased as:

(42) In wounding Caesar, Brutus caused Caesar to die.

In this example, it is the object of the verb that is involved in the resulting event, i.e. Caesar died. Cresswell (1981) showed that in some cases it is the subject of the verb that is involved in the resulting event:

(43) Catherine *fatally* slipped.

The adjective *fatal* also has a causal meaning:

(44) Caesar's wound was *fatal*.
(45) Guinevere's *fatal* walk ...

Other examples of causal adverbs cited by Cresswell are

- the adverbs of perception, e.g. *audibly*, *visibly*.
- other adverbs that are marginally perceptual, e.g. *manifestly*, *patently*, *publicly*, *conspicuously*.
- adverbs that involve the notion of a result whose properties are context dependent, e.g. *successfully*, *plausibly*, *conveniently*, *amusingly*, *pleasantly*.
- adverbs that suggest tendencies, liabilities, disposition or potencies, e.g. *irrevocably*, *tenuously*, *precariously*, *rudely*.
- adverbs that refer not to causes but to effects, e.g. *obediently*, *gratefully*, *consequently*, *painfully*.
- adverbs of means, e.g. *mechanically*, *magically*.

This study did not make use of causal adverbs and adjectives, except for those covered in Altenberg's Typology described in Section 4.2. Causal adverbs and adjectives are not well studied, and a comprehensive list of such adverbs and adjectives has also not been identified.

3.7. Implicit Causal Attribution by Verbs

Some verbs have "causal valence." They tend to assign causal status to their subject or object. Some verbs give the reader the impression that the cause of the event is the participant occupying the syntactic subject position of the sentence. Other verbs suggest that the cause of the event is the participant in the object position. This phenomenon has been referred to as the *implicit* or *inherent causality* property of verbs (Brown & Fish, 1983; Caramazza, Grober, Garvey & Yates, 1977).

This implicit causal attribution can be made explicit by requiring the reader to determine whether an ambiguous anaphoric pronoun refers to the subject or object of the verb. Garvey and Caramazza (1974) had subjects complete sentences of the form "NP Verb NP because Pronoun . . .", for example:

(46) The mother punished her daughter because she ___

In completing the sentence, the subject automatically makes a choice as to whether "she" refers to mother or daughter. Garvey and Caramazza also asked subjects to supply responses to questions of the form

(47) Why did the director criticize the actor?

In constructing a response to this question the subject decides whether the reason lies with the director or with the actor.

Garvey and Caramazza (1974) found that for the verbs *confess, join, sell, telephone, chase,* and *approach*, subjects tended to assign the pronoun to the subject of the verb, for example,

(48) The prisoner confessed to the guard because he <u>wanted to be released.</u>

Subjects also tended to respond to the question by assigning the reason to the subject of the verb. For the verbs *kill, fear, criticize, blame, punish, scold, praise, congratulate,* and *admire*, subjects tended to assign the pronoun and the reason for the event to the object of the verb, for example,

(49) The mother punished her daughter because she <u>broke an antique vase.</u>

For the verbs *help, recognize, give, argue with,* and *miss*, the subjects did not agree in assigning the pronoun and the reason for the event. Assignment to subject and to object occurred equally often, for example,

(50) Why did John give Walt the book?

Because he didn't need it anymore.
Because he wanted to read it.

The implicit causal attribution effect of verbs has been replicated using other experimental methods (Brown & Fish, 1983; Caramazza, Grober, Garvey & Yates, 1977).

Garvey, Caramazza and Yates (1974/1975) found that the causal bias of verbs is not discrete but is a continuum. Verbs vary in the degree of bias in attributing causality to the subject or object. When the attribution bias for a verb is defined as the percentage of subjects who complete the sentences of the form "NP Verb NP because Pronoun . . ." by assigning the pronoun to the subject, bias values range continuously from 0.0 to 1.0.

Garvey, Caramazza and Yates (1974/1975) found that the implicit causal attribution of a verb can be modified or reversed by the following:

- Negating the verb. The sentence "the doctor *did not* blame the intern . . ." produced a smaller attribution bias towards the object than when the verb is not negated.
- Converting the sentence to passive voice. When sentences are passivized, there is a shift in the direction of causal attribution towards the surface structure subject of the verb.
- Changing the nouns occupying the subject and object position. Garvey *et al.* (1974/1975) suggested that the relative social status of the participants occupying the subject and object position influence the causal attribution. For the sentence,
  (51) The father praised his son . . .
  causality tends to be imputed more to the object than for the sentence
  (52) The son praised his father . . .

To be able to make use of the implicit causality feature of verbs in automatic text processing, it is important to be able to identify classes of verbs that tend to attribute causality in one direction or the other. It is obviously not feasible to determine empirically the direction of implicit causal attribution for every verb.

One class of verbs that follow a clear rule for attributing causality is the class of experiential verbs, which describes someone having a particular psychological or mental experience. Experiential verbs such as *like* and *fear* assign the thematic role of experiencer (the role of having a given experience) to the subject of the verb indicating that it is the subject who has the experience. Experiential verbs like *charm* and *frighten* assign the *experiencer* role to the object of the verb. Several studies have confirmed that experiential verbs attribute causality to the stimulus (the entity or participant giving rise to a certain experience) regardless of whether the stimulus occupies the syntactic subject or object position and whether the sentence is active or passive (Au, 1986; Brown & Fish, 1983; Caramazza, Grober, Garvey & Yates, 1977; Corrigan, 1988). Here are some examples:

(53a) John (experiencer) fears Bill (stimulus). (Cause is attributed to *Bill*)
(53b) John (stimulus) frightens Bill (experiencer). (Cause is attributed to *John*.)

(54a) John (experiencer) likes Bill (stimulus). (Cause is attributed to *Bill*.)
(54b) John (stimulus) charms Bill (experiencer). (Cause is attributed to *John*.)

Corrigan (1988) found that action verbs that have derived adjectives referring to the actor (i.e. subject) assign greater causal weight to the subject. She referred to these action verbs as *actor verbs*. Some examples are given below:

Actor verbs	Derived adjectives referring to the actor
defy	defiant
help	helpful
dominate	domineering
criticize	critical

For the sentence "John defies Bill," the derived adjective *defiant* refers to the subject *John*. Corrigan found that *actor verbs* tend to attribute causality to the subject regardless of whether the subject and/or object are animate. On the other hand, non-actor verbs (that either have no derived adjectives, or have derived adjectives that refer to the object) tend to attribute causality to the object if the subject is animate and has greater status than the object. Non-actor verbs tend to attribute causality to the subject if the subject is inanimate or if both subject and object are animate and have neutral status (i.e. names of persons for which the status is not known).

Implicit causal attribution of verbs was not used in this study to identify causal relations. The implicit causality feature of verbs is easily attenuated, nullified and even reversed by other factors. The *implicit* causal attribution suggests rather than explicitly expresses a causal relation, and so does not have the same status as the other indications of causal relation described in this chapter.

3.8. Multiple Meanings and Functions of Words

Based on the types of causal expressions described in this chapter, I constructed a set of linguistic patterns that could be used by a computer program to identify causal relations in text. The features of the linguistic patterns are described in the next chapter. The effectiveness of the linguistic patterns for identifying causal relations is limited by how reliably the causal words and constructions described in this chapter do indicate a causal relation.

Many of the words used for indicating causal relations have other functions and meanings. The adverbial *since* can be used to introduce not only a causal clause but also a temporal clause. *As* can suggest not only cause but also time (synonymous with *while*) or imply degree or manner. Causal verbs like *break* have other non-causative meanings:

(55a) The plane *broke* the sound barrier.
(55b) We *broke* our journey at Rochester.
(55c) They *broke* the law.

Schleppegrell (1991) showed that the word *because* has other functions than as a subordinating conjunction with causal meaning. One of these functions is as a "discourse-reflexive" link which introduces a reason why the speaker knows or has asserted a proposition. The following is an example from Rutherford (1970):

(56) He's not coming to class, because he just called from San Diego.

Here, the clause "he just called from San Diego" does not express a reason why "he's not coming to class," but rather expresses *the speaker's reason for knowing* or asserting this proposition. One can, of course, argue that sentence (56) is really an abbreviated form of:

(57) I know he's not coming to class, because he just called from San Diego.

If-then conditionals can also have this discourse-reflexive function:

(58) If John spent the night with Mary, then he didn't commit the murder.
(59) If John says that he spent the night with Mary then he's a liar.

In these examples, the antecedent (the *if* part of the sentence), rather than causing the consequent, introduces a potential reason for the speaker to believe the statement in the consequent (the *then* part).

Because words often have several possible meanings and functions, the reader often has to make use of general knowledge and contextual information to determine whether the words are used to indicate a causal relation or something else. Since the linguistic patterns constructed in this study do not make use of general knowledge and contextual information, it is inevitable that the linguistic patterns will erroneously identify some sentences as containing a causal relation.

3.9. Summary

In this chapter, I have described the following ways in which causal relations are expressed in text:

- by using causal links to link two phrases, clauses or sentences, thus indicating a causal relation between them.
- by using causal verbs -- verbs that have a causal meaning. In this study, verb entries in the *Longman Dictionary of Contemporary English* (2nd ed.) were scanned to identify such causal verbs.
- by using resultative constructions.
- by using "if ... then ..." constructions.
- by using causal adverbs and adjectives.
- by using verbs that implicitly attribute cause to the subject or object of the verb.

In this study, linguistic patterns were constructed to be used by a computer program to identify sentences containing causal links, causal verbs, adjectival resultative phrases or if-then constructions. The phrases, clauses or sentences connected by the causal link, causal verb, resultative construction or if-then conditional were then extracted as the cause and effect.

Causal adverbs and adjectives were not used in this study to identify cause and effect because of the difficulty in identifying these adverbs and adjectives. Also not used was the implicit causal attribution of verbs. Such causal attribution is only implied and not explicitly expressed. Moreover, there is no automatic procedure for determining the direction of the causal attribution for every verb.

This study limited itself to causal relations that are explicitly indicated using causal links, causal verbs, resultatives and if-then conditionals. However, because words have multiple meanings and functions, accurate identification of causal relation requires domain knowledge and contextual information. Domain knowledge and contextual information were not used in this study, and this resulted in some sentences being erroneously identified as containing causal relations.

The next chapter describes the features of the linguistic patterns constructed, and evaluates how effective the linguistic patterns are in identifying causal relations in Wall Street Journal articles.

CHAPTER 4
IDENTIFICATION OF CAUSAL RELATIONS IN TEXT

4.1. Introduction

In this study, linguistic patterns were constructed and subsequently used by a computer program to locate causal relations in natural language text, and to extract the cause-effect information from the text. This chapter describes the features of the linguistic patterns, how they were constructed and how they were used to identify and extract cause-effect information in text. The effectiveness of the linguistic patterns for identifying and extracting cause-effect information is evaluated to address the first research question:

> **Research question 1**
> How effectively can cause-effect information expressed in sentences be identified and extracted using linguistic patterns?

4.2. Description of the Linguistic Patterns

The linguistic patterns constructed in this study specify various ways in which the causal relation can be expressed in the English language. To identify causal relations in a document, a computer program locates all parts of the document that match with any of the linguistic patterns. "Slots" in a linguistic pattern indicate which part of the text is the *cause* and which the *effect*. For example, the pattern

[effect] *is the result of* [cause]

indicates that the part of the sentence following the phrase "is the result of" represents the *cause* and the part of the sentence preceding the phrase represents the *effect*.

Three sets of patterns are constructed in this study:

1. patterns involving a causal link or an *if-then* conditional that links two phrases within a sentence,
2. patterns involving a causal link that links two adjacent sentences,
3. patterns involving causal verbs and resultative constructions.

Causal links, *if-then* conditionals, causal verbs and resultatives were all described in Chapter 3. The first and third set of patterns were constructed to identify causal relations that occur *within a sentence*, while the second set of patterns was constructed to identify causal relations *between adjacent sentences*. The three sets of patterns are listed in Appendix 4.

Each pattern consists of a sequence of tokens separated by a space. Each token indicates one of the following:

- a particular word
- a word having a particular part-of-speech label (e.g. an adjective)
- a particular type of phrase (e.g. noun phrase)
- a set of subpatterns (as defined in a *subpatterns file*)
- any verb from a particular group of verbs (as defined in a *verb groups file*)
- a slot to be filled by one or more words representing the cause or the effect
- any word or phrase (i.e. a wild card symbol).

NO.	RELATION	PATTERN
(1)	C	[1] &AND because of &THIS[1],[2] &._ Example: It was raining heavily and because of this the car failed to brake in time.
(2)	-	&NOT because Example: It was not because of the heavy rain that the car failed to brake in time.
(3)	C	it &AUX &ADV_ because of [1] that [2] &._ Example: It was because of the heavy rain that the car failed to brake in time.
(4)	C	it &AUX &ADV_ because [1] that [2] &._ Example: It was because the rain was so heavy that the car failed to brake in time.
(5)	C	&C2_ &[2](AND_THIS) &AUX &ADV_ because of [1] &._ Example: The car failed to brake in time and this was because of the heavy rain.
(6)	C	&C2_ &[2](AND_THIS) &AUX &ADV_ because [1] &._ Example: The car failed to brake in time and this was because it was raining heavily.
(7)	C	&C because of &[N:1],[2] Example: John said that because of the heavy rain , the car failed to brake in time.
(8)	C	&C2_ [2] because of [1] &._ Example: The car failed to brake in time because of the heavy rain.
(9)	C	&C because &[C:1],[2] Example: Because it was raining so heavily , the car failed to brake in time.
(10)	C	&C2_ [2] because [1] &._ Example: The car failed to brake in time because it was raining so heavily.

Note: "C" in the second column indicates that the pattern can be used to identify a cause-effect relation. The symbol "-" in the second column indicates a null relation, i.e. the pattern does not identify the presence of any relation.

Figure 2. Examples of linguistic patterns for identifying the cause-effect relation.

2 gives, as examples, some of the patterns involving the word *because*. The symbols used in the patterns in 2 are explained below.

[1] and *[2]* in the patterns represent slots to be filled by the first and second member of the relation respectively, the first member of the causal relation being the cause and the second member the effect. The type of phrase or word that may fill a slot may also be indicated. The symbol *[N:1]* indicates that the slot for *cause* is to be filled by a noun phrase, whereas *[n:1]* indicates that the slot is to be filled by a noun. *[C:1]* indicates that the slot is to be filled by a clause.

The symbol *&* followed by a label in uppercase refers to a set of subpatterns (usually a set of synonymous words or phrases). For example, *&AUX* in patterns (3) to (6) of 2 refers to auxiliary verbs like *will*, *may*, and *may have been*. *&C* and *&C2_* in patterns (5) to (10) refer to subpatterns that indicate the beginning of a clause. *&._* refers to a set of subpatterns that indicate the end of a clause or sentence, and this of course includes the period. *&[2](AND_THIS)* in patterns (5) and (6) refers to the following set of three subpatterns:

[2] &AND &THIS/IT
[2] &AND &THIS [1]
[2]

The first two subpatterns above contain the tokens *&AND*, *&THIS/IT* and *&THIS*, each referring to a set of subpatterns. The example illustrates that a subpattern can contain tokens that refer to a set of subpatterns.

The sets of subpatterns are defined in a *subpatterns file*. Each set of patterns is associated with a particular *subpatterns file*. The contents of the *subpatterns files* are listed in Appendix 4. Appendix 4 also describes in greater detail the features of the linguistic patterns and the notation used.

For each set of patterns, the patterns are tried in the order listed in the set. Once a pattern is found to match a sentence (or some part of a sentence), all the words that match the pattern (except for the words filling the slots) are flagged, and these flagged words are not permitted to match with tokens in any subsequent pattern, except for the tokens that represent slots. In other words, the flagged words are permitted to fill slots in subsequent patterns, but are not allowed to match with any of the other tokens. So, the order in which patterns are listed in the set is important. As a rule, a more "specific" pattern is listed before a more "general" pattern. A pattern is more specific than another if it contains all the tokens in the other pattern as well as additional tokens not in the other pattern.

Consider the following three patterns:

(1) [1] and because of this , [2]
(2) because [1] , [2]
(3) [2] because [1]

Pattern (1) is more specific than patterns (2) and (3), and pattern (2) is more specific than pattern (3). All the sentences that pattern (1) will match, patterns (2) and (3) will match also. For example, all three patterns will match the following sentence:

(4) It was raining heavily and because of this, the car failed to brake in time.

Note that a pattern does not need to match the whole sentence for a match to occur. A pattern needs to match just some part of the sentence for a causal relation to be identified. So, pattern (2) does not require the word *because* to appear at the beginning of the sentence. Pattern (2) will find a match in sentence (4) because there is a comma after the word *because* in sentence (4).

However, only pattern (3) will match the sentence:

(5) The car failed to brake in time because it was raining heavily.

Pattern (1) will not match the sentence because the sentence does not contain the phrase *and because of this*. Pattern (2) will not match sentence (5) because pattern (2) requires that there be a comma after the word *because*. So, pattern (3) is more general than patterns (1) and (2) in the sense that pattern (3) contains fewer constraints.

Although all three patterns will match sentence (4), only pattern (1) will correctly identify the cause and the effect in the sentence. Applying pattern (1) to sentence (4), we obtain:

cause: it was raining heavily
effect: the car failed to brake in time

Applying, pattern (2) to sentence (4), we obtain:

cause: of this
effect: the car failed to brake in time

which, although not wrong, is not as informative as the result of applying pattern (1). On the hand, applying pattern (3) to sentence (4) yields the incorrect result:

cause: of this, the car failed to brake in time.
effect: it was raining heavily and

Because pattern (1) is listed before patterns (2) and (3), pattern (1) will be applied to the sentence first and the words *and because of this* are flagged in the sentence so that they are not permitted to match with any of the non-slot tokens in patterns (2) and (3).[23] In particular, the word *because* is flagged and is not permitted to match with the token *because* in patterns (2) and (3).

In 2, pattern (2) is associated with a "-" symbol in the *relation* column (second column) indicating that the pattern "&NOT because" does not identify the presence of any relation. If a sentence contains the words *not because*, the words will match with this pattern and will be flagged so that they will not match with tokens in any other pattern. The effect of this pattern is that if the word *because* is preceded by a negation, no pattern will be able to use the word to identify the presence of a causal relation.

Negated causal relation is ignored in this study. This is because hardly any of the query statements used in this study contain a negated causal relation. Negation is a difficult problem in information retrieval, and it is not addressed in this study.

4.3. Prior Text Processing Assumed by the Linguistic Patterns

In order to apply the linguistic patterns to extract cause-effect information, the text has to be pre-processed in the following ways:

- the beginning and end of sentences have to be identified, and each sentence placed on a separate line
- words in the text have to be tagged with part-of-speech labels

[23] Punctuation marks are not flagged when they match with a token in a pattern. This is because a punctuation mark does not have a meaning the way a word has. Punctuation marks only help to indicate the syntactic structure of the sentence. In the linguistic patterns constructed in this study, punctuation marks are used not so much to identify causal relations as to identify where the cause or effect phrase begins or ends in the sentence. It is necessary to use punctuation marks in the patterns only because sentences are not parsed in this study. Flagging punctuation marks will result in some errors because patterns are then prevented from "using" a punctuation mark once it has matched with a token in an earlier pattern. I make use of the period "." in many of the patterns to locate the end of the sentence. If the period is flagged once a pattern is found to match the sentence, this will prevent most other patterns (i.e. those containing a period) from matching the sentence.

- the boundaries of phrases (e.g., noun phrases) have to be marked with brackets.

Sentence and phrase boundary identification was done using text processing programs developed in the DR-LINK project (Liddy & Myaeng, 1993; Liddy & Myaeng, 1994). The phrase boundary bracketer developed in the DR-LINK project was used to identify noun phrases, prepositional phrases, past participle phrases, present participle phrases (i.e. non-finite clauses introduced by an *-ing* verb), as well as clauses (Myaeng, Khoo, & Li, 1994). Part-of-speech tagging was performed using the POST tagger (obtained from BBN Systems and Technologies) which uses 36 part-of-speech tags (Meteer, Schwartz, & Weischedel, 1991).

The part-of-speech and phrase labels used are given in Appendix 2, together with a sample of processed text.

4.4. Procedure Used in Constructing the Linguistic Patterns

The linguistic patterns were constructed through several cycles of inductive and deductive steps. The *inductive step* involves examining sample sentences that contain a causal relation, and deciding what features in the sentences indicate the causal relation. Linguistic patterns are then constructed that specify these features. In other words, from an example of a sentence containing a causal relation, I hypothesize that other sentences sharing certain features with this sentence also contain a causal relation. For example, given the example sentence

(6) The car failed to brake in time because it was raining heavily.

one might construct the pattern

(7) [2] because [1]

specifying that any sentence containing the word *because* contains a causal relation, and that the words preceding *because* represents the effect and the words following *because* represents the cause.

The *deductive step* involves applying the patterns that have been constructed to additional sample sentences, using the patterns to identify cause and effect in these sentences. The result of applying the patterns is examined. The errors can be divided into three types:

1. *Misses*. This refers to causal relations in the sample sentences that are missed by the patterns. This kind of error reduces the *recall measure*, one of the measures used in this study to evaluate the effectiveness of the patterns.
2. *False hits*. These are instances where the patterns incorrectly identify a causal relation in a sentence when in fact there is no causal relation in the sentence. Such errors reduce the *precision measure*, another evaluation measure used in this study.
3. *Syntax errors*. These are instances where the patterns correctly find a causal relation in a sentence, but the wrong parts of the sentence are identified as the cause and the effect. These errors reduce both the recall and precision measures.

Based on these errors, another inductive step can be applied to modify the set of patterns and improve its effectiveness. How the set of patterns should be modified depends on the type of error.

Misses can be handled in two ways. Additional patterns can be constructed to identify the causal relations that are missed by the current set of patterns. Alternatively, one of the current patterns can be modified and made more "general" by leaving out some features from the pattern so that the pattern will now match the sentence and identify the causal relation previously missed.

False hits can be handled in two ways. One or more patterns can be made more specific by adding features to the patterns so that the new set of patterns will no longer match the sentence and will not identify a causal relation in the sentence. Alternatively, a pattern can be constructed to match the sentence and be assigned the null relation. Pattern (2) in 2 is an example of a pattern associated with a null relation. The purpose of a pattern with a null relation is to exclude certain phrases from being identified as containing a causal relation. As an example, the word *since* can mean *because* as in the following example:

(8) It took a longer time for the car to stop *since* the road was so slippery.

Since can also have the meaning "after a certain time" as in:

(9) I've not seen her *since* January.
(10) I've had this cold *since* before Christmas.
(11) It has been a long time *since* I visited my parents.

The pattern "[2] since [1]" would correctly identify sentence (8) as containing a causal relation. It would, however, incorrectly identify sentences (9) to (11) as containing a causal relation. To prevent this, the following patterns associated with the null relation can be added to the set of patterns and placed before the pattern "[2] since [1]":

Relation	Pattern
-	since &TIME
-	since before
-	time since

(Note: "&TIME" refers to a set of words related to time and duration, including the months of the year.)

Syntax errors can be handled by adding a more specific pattern before the pattern that matched the sentence. For example, the pattern "[2] because [1]" will not correctly identify the cause and effect in the following sentence:

(12) It was raining heavily and because of this, the car failed to brake in time.

This can be remedied by adding the pattern

[1] and because of this , [2]

before the pattern

[2] because [1]

This cycle of deduction and induction is repeated until few errors are made by the set of patterns.

The linguistic patterns in this study were initially constructed based on:

1. a review of the linguistics literature as given in Chapter 3. Altenberg's (1984) list of words and phrases that indicate causal linkage was particularly useful. His list was compiled from many sources, including Greenbaum (1969), Halliday and Hasan (1976) and Quirk, Greenbaum, Leech and Svartvik (1972).
2. a set of linguistic patterns for identifying causal relations that I developed in an earlier study (Venkatesh, Myaeng, & Khoo, 1994). This set of patterns was developed based on a sample

of 60 sentences containing causal relations which were taken from the ABI/Inform database, a database of abstracts of articles and books from the management literature.

3. the list of causal verbs given in Appendix 3. I have described in Chapter 3 (Section 3.3) how the list of causal verbs was constructed.

The initial set of patterns was applied deductively to a sample of 10 documents from Wall Street Journal. Based on the errors found, the set of patterns was modified so that most of the errors for this sample were eliminated. The modified set of patterns was applied to a second sample of 10 documents and the cycle of deduction and induction was repeated. Altogether 50 consecutive documents from Wall Street Journal were used.

The effectiveness of the patterns for identifying cause and effect depends to some extent on the number of examples used to develop the patterns. Some of the words that can indicate a causal relation did not occur often in the 50 sample documents, and I was not confident that the set of patterns would accurately identify cause and effect in sentences containing these words. The following words were identified for further investigation:

> advantage, as, because, by, cause, critical, disadvantage, effect, factor, for, if, reason, require, responsible, since, so, through, to

For each of the above words, sample sentences containing the word were extracted from the document collection, and several cycles of deduction and induction were carried out. No fixed number of sample sentences were used. For each of the words, the cycle of deduction and induction was iterated until further adjustments to the set of patterns yielded no improvement in the effectiveness of the patterns or there was no obvious way to improve the patterns.

Finally, the set of patterns was applied to a sample of 200 pairs of sentences randomly selected from four months of Wall Street Journal documents. Final adjustments to the set of patterns were made based on the errors found.

4.5. Evaluation of the Effectiveness of the Patterns in Identifying Cause and Effect in Sentences

The evaluation is based on a random sample of 509 pairs of adjacent sentences and 64 single sentences (1082 sentences in all) taken from about four months of Wall Street Journal articles.[24] The effectiveness of the computer program in identifying and extracting cause-effect information from Wall Street Journal using the patterns is evaluated by comparing the output of the computer program against the judgments of two human judges, who were asked to identify causal relations in the sample sentences. One judge (identified as judge A) was a professional librarian. Though English was not her native language, it was effectively her first language, having been educated entirely in English. The second judge (identified as judge B) was a native speaker of English and was a graduate student in the School of Education at Syracuse University.

The judges were "trained" using a training set of 200 pairs of sentences randomly selected from Wall Street Journal. The judges were asked to identify causal relations in the training set of sentences, and their judgments were compared with the causal relations that I had identified in the sentences. I then discussed with each judge the instances where their judgments differed from mine. Difficulties encountered by the judges were noted (discussed in a later section) and the

[24] The sampling program sought to extract only pairs of adjacent sentences, but had to make do with single sentences when there was only one sentence in the article or if the sentence was a headline. Also, some lines in Wall Street Journal are not sentences (they may be a statistic or the name of the contributing writer) and such lines had to be dropped from the sample.

instructions to judges were modified. The modified instructions to the judges are given in Appendix 1.

The evaluation is divided into two parts. Part 1 of the evaluation focuses on whether the computer program can identify the presence of a causal relation and the direction of the causal relation. Part 2 evaluates how well the computer program can identify the "scope" of the causal relation, i.e. can correctly extract all the words in the text that represent the cause and all the words that represent the effect. Since a *cause* and *effect* can comprise more than one word, there will be instances where the computer program extracts more words or fewer words than is appropriate. For example, given the sentence

The surgeon general said that cigarette smoking is likely to cause lung cancer.

the computer program might correctly extract "lung cancer" as the effect but incorrectly extract "the surgeon general said that cigarette smoking is likely" as the cause. In extracting the cause, the computer program has extracted more words than it should have. In part 1 of the evaluation, I consider that the computer program has correctly identified the presence of a causal relation if the program manages to extract some part of the cause and some part of the effect. In part 2, I focus on the number of the words that are correctly extracted as the cause and effect.

4.5.1. Evaluation part 1: identifying the presence of a causal relation

Causal relations identified by 2 human judges

Number of causal relations identified by judge A: 615
(607 are within a sentence, and 8 are across two sentences)

Number of causal relations identified by judge B: 174
(172 are within a sentence, and 2 are across two sentences)

Number of causal relations identified by both A and B (intersection of judgments A and B): 161
(63 involve causal links, and 98 involve causal verbs)

Number of causal relations identified by either A or B (union of judgments A and B): 628

Causal relations identified by computer program

Total number of causal relations identified by computer program: 437
Number involving a causal link: 117
(Only 2 relations are between sentences)
Number involving a causal verb: 320

Comparison between human judgments and judgments by computer program

Number of causal relations identified by both computer program and judge A: 279
Number involving causal links: 86
Number involving causal verbs: 193

Number of causal relations identified by both computer program and judge B: 110
Number involving causal links: 49
Number involving causal verbs: 61

Number of causal relations identified by computer program and both human judges: 109
Number involving causal links: 49
Number involving causal verbs: 60

Table 1. Number of causal relations identified by the computer program and the human judges.

The performance measures used are *recall* and *precision*. *Recall*, in this context, is the proportion of the causal relations identified by the human judges that are also identified by the computer program. *Precision* is the proportion of causal relations identified by the computer program that are also identified by the human judges. *Recall* measures how comprehensive the identification of causal relations is, whereas *precision* measures what proportion of the causal relations identified by the computer program is in fact correct.

Two human judges (identified as judge A and judge B) were asked to identify causal relations in the sample of 1082 sentences. The results are given in 1. I shall highlight the more important results.

Judge A identified many more causal relations than judge B (615 for judge A and 174 for judge B). 91% of the causal relations identified by B were also identified by A, and 26% of the causal relations identified by A were also identified by B. Clearly, the causal relations identified by judge B were largely a subset of the relations identified by judge A. Judge A and judge B had 161 causal relations in common. I shall refer to the causal relations identified by both A and B as the intersection set, and the causal relations identified by either A or B as the union set.

Why was there such a big difference in the number of causal relations identified by judge A and judge B? The fact that most of the causal relations picked out by judge B were also identified by judge A indicates a high degree of consistency between the two judgments. It is just that judge A picked out a lot more causal relations. Judge A spent much more time on the task than judge B (about three or four times more) and went over the sample sentences a few times. So, judge A's judgments were more thorough and probably more liberal than B's. Judge B's list of causal relations probably represents the more obvious causal relations.

In calculating *recall*, I compared the judgments made by the computer program with the intersection set, which was made up of causal relations identified by both human judges. There is some amount of subjectivity involved in identifying causal relations in text -- especially in deciding whether the causal relation is explicitly expressed or merely implied. Taking the intersection set of two judgments eliminates idiosyncratic judgments by either judge, and ensures that the causal relations used to evaluate the effectiveness of the computer program are those that are clearly expressed in the text. The intersection set probably also represents the more obvious causal relations. Of the causal relations in the intersection set, 109 were picked up by the computer program, giving a recall of 68% (109/161). The width of the 95% confidence interval was ±7%. Of the causal relations in the intersection set, 63 involved causal links and 98 involved causal verbs. For causal links the recall was 78% (49/63), whereas for causal verbs the recall was 61% (60/98).

In calculating *precision*, I compared the judgments made by the computer program with the union set of the two human judgments. The purpose of calculating *precision* was to find out how many of the causal relations identified by the computer program were *reasonable*. I assumed that a decision made by either judge was reasonable. 280 of the causal relations identified by the computer program were in the union set, giving a precision of 64% (280/437). The width of the 95% confidence interval was ±5%. For causal links the precision was 74% (86/117), whereas for causal verbs the precision was 61% (194/320).

Not all the causal relations identified by the computer program but not by the judges were clearly wrong. In reviewing the causal relations identified by the computer, I found that 33 of the relations not picked out by the judges might be considered to be correct. If we give the computer program the benefit of the doubt when the causal relation extracted by the program was not clearly wrong, the precision was 72% (313/437) -- 79% (92/117) for causal links and 69% (221/320) for causal verbs.

Identifying the scope of the cause

For causal links (averaged over 86 causal relations):

Average recall =	0.98
Average precision =	0.96

For causal verbs (averaged over 194 causal relations):

Average recall =	0.93
Average precision =	0.94

Identifying the scope of the effect

For causal links (averaged over 86 causal relations)

Average recall =	0.96
Average precision =	0.91

For causal verbs (averaged over 194 causal relations)

Average recall =	0.86
Average precision =	0.98

Table 2. Evaluation of how accurately the computer program can identify the scope of the cause and the effect.

4.5.2. Evaluation part 2: determining the scope of the causal relation

This section evaluates how accurately the computer program can determine what part of the text is the cause and what part is the effect. For this evaluation, I examined each causal relation that was identified by the computer program as well as by either of the human judges. (In other words, this evaluation is done using only those instances where the computer program correctly identified the *presence* of a causal relation.) I compared the words that were extracted by the computer program as the *cause* with the words that were identified by a human judge as the *cause*, and calculated the measures of recall and precision -- *recall* being the proportion of words extracted by the human judge that were also extracted by the computer program, and *precision* being the proportion of words extracted by the computer program that were also extracted by the human judge. The recall and precision measures were also calculated for the *effect* part of the relation. The recall and precision figures were then averaged across all the causal relations. The results are given in 2.

For the *cause* part of the relation, the average recall was 98% for causal links and 93% for causal verbs. The average precision was 96% for causal links and 94% for causal verbs. For the *effect* part, the average recall was 96% for causal links and 86% for causal verbs. The average precision was 91% for causal links and 98% for causal verbs.

4.6. Sources of error

A. Errors of commission

No. of instances identified by the computer program to be a causal relation involving a causal link, but not identified by either of the human judges: 31

Reasons why errors occurred:

A1. No. of these instances that, in my opinion, can be considered to be correct: 6
A2. Unexpected sentence structure resulting in the wrong part of the sentence extracted as the cause or the effect: 1
A3. Unexpected sentence structure resulting in the program identifying a causal relation where there is none: 2
A4. Linking words not used in a causal sense: 22

B. Errors of omission

No. of causal relations *not* identified by the program: 14

Reasons why errors occurred:

B1. Unexpected sentence structure resulting in the causal relation not picked up by the system: 2
B2. Unexpected sentence structure resulting in the wrong part of the sentence extracted as the cause or the effect: 1
B3. Causal link is not in the list of patterns: 11

Table 3. Analysis of errors made by computer program in identifying causal relations involving causal links

The errors made by the computer program in identifying causal relations were examined to see why the errors occurred. I shall divide the discussion into:

1. errors involving causal links
2. errors involving causal verbs.

For each of these, I shall discuss both errors of commission (instances where the computer program indicated there was a causal relation when in fact there wasn't) and errors of omission (causal relations that the computer program failed to identify).

4.6.1. Errors involving causal links

The reasons for the errors involving causal links are summarized in 3. Most of the errors of commission were due to the fact that the same words and sentence constructions that are used to indicate cause-effect can be used to indicate other kinds of relations as well.

The sentence pattern that gave rise to the highest number of errors of commission was the pattern

> [effect] by [present participle phrase: cause]

which accounted for 7 errors in the sample sentences. This pattern was constructed to identify causal relations in sentences like:

(13) [$_{\text{effect}}$ Japan has become a major economic power] mainly by [$_{\text{cause}}$ exporting to the U.S.]

However, this sentence construction can also be used to indicate the manner in which something is done, as in the following examples:

(14) Secretary Baker has done a service just *by* mentioning the word in public.

(15) Senator Proxmire challenged the nominee *by* disputing economic forecasts he had made during the Ford administration.

In sentence (14), "mentioning the word in public" was how Secretary Baker did a service, not why he did it. Similarly, in sentence (15), "disputing economic forecasts ..." was the manner Senator Proxmire challenged the nominee, rather than the reason he challenged the nominee. The pattern correctly identified 4 causal relations in the sample sentences, but incorrectly identified a causal relation in 7 other sentences.

The conjunction "as" accounted for 4 of the errors of commission, and "if .. then" constructions accounted for 3 errors.

I turn now to errors of omission. Most of the errors of omission were due to particular kinds of linking words or sentence constructions not included in my list of patterns. Many of these linking words and sentence constructions are seldom used for indicating cause and effect. Below are 4 of the sentences that contain causal relations not picked up by the computer program:

(16) [$_{\text{effect}}$ Crop conditions improved considerably in several states] *with* [$_{\text{cause}}$ widespread rains in June.]

(17) It's such a volatile stock -- [$_{\text{cause}}$ the slightest thing goes wrong] *and* [$_{\text{effect}}$ the stock takes a nosedive.]

(18) Under [$_{\text{cause}}$ special federal rules,] [$_{\text{effect}}$ Anadarko recently has been able to sell some of the gas dedicated to Panhandle on the spot market instead.]

(19) [$_{\text{effect}}$ Northop will have $550 million in revenue this year] *from* [$_{\text{cause}}$ electronic "countermeasures" equipment such as aircraft jammers.]

For the above sentences, inferencing from general knowledge is needed to identify the causal relations.

4.6.2. Errors involving causal verbs

C. Errors of commission

No. of instances identified by the computer program to be a causal relation involving a causal verb, but not identified by either of the human judges: 126

Reasons why errors occurred:

C1. No. of instances that can be considered to be correct: 27
C2. Error in part-of-speech tagging (a word is incorrectly tagged as verb): 4
C3. The noun phrase occupying the object position is not the "patient" of the verb: 8
C4. Unexpected sentence structure resulting in the wrong part of the sentence extracted as the cause or the effect: 15
C5. Unexpected sentence structure where the cause is not specified in the sentence: 10
C6. The sentence having the syntactic pattern *V-NP-Adj* is not a resultative sentence: 4
C7. The verb is not used in its causal sense: 58

D. Errors of omission

No. of causal verbs identified by both judges but not identified by program: 38

Reasons why errors occurred:

D1. Error in part-of-speech tagging: 3
D2. Error in phrase bracketing: 5
D3. Unexpected sentence structure resulting in the causal relation not picked up by the program: 13
D4. Causal verb is used in nominalized form: 2
D5. Resultative construction not handled: 1
D6. Verb is not in my list of causal verbs: 14

6 of the verbs involve an unusual sense of the verb.

8 of the verbs can, arguably, be included in the list of causal verbs. (The 8 verbs are: benefit, bolster, design, drive down, require, credit (somebody) for, highlight, and ban (somebody) from.)

Table 4. Analysis of errors make by computer program in identifying causal relations involving causal verbs

The reasons for the errors in identifying causal relations involving causal verbs are summarized in 4. Some of the reasons listed in the table require an explanation.

Reason C3 refers to sentences such as the following:

(20) Forest products segment sales increased 11.6% to $157.6 million.
(21) The December-to-March premium increased 0.15 cent to 7.9 cents.
(22) Treasury bond futures for December settlement dropped almost two points to 78 21/32 at the Chicago Board of Trade.
(23) Kawasaki Steel dropped 7 to 333.
(24) It moved up 3/8 to 24 7/8.

In each of the above sentences, the noun phrase following the verb is not assigned a "patient" role by the verb, i.e. the noun phrase does not refer to the object affected by the action denoted by the verb. Rather, the noun phrase indicates the magnitude of the process denoted by the verb. For example, in sentence (20), "11.6%" was not the thing that increased. It was the subject of the verb, "forest products segment sales", that increased, and "11.6%" only specified the magnitude of the increase. It is not difficult to write a program to check whether the noun phrase following a verb denotes a magnitude. However, it is difficult to avoid errors completely. For example, it is not easy to distinguish between sentence (a) and sentence (b) in the following pairs of sentences:

(25a) Jim Grocer dropped 2 eggs.
(25b) The Dow Jones Industrial Average dropped 12 points.

(26a) Jim Grocer tried to carry a dozen eggs but dropped 2.
(26b) Kawasaki Steel dropped 7.

(27a) Jim Grocer dropped 20 cents.
(27b) The December-to-March premium dropped 20 cents.

In the sentences labeled (a), the noun phrase following the verb is the "patient" of the verb. In the sentences labeled (b), the noun phrase following the verb specifies the magnitude of the process denoted by the verb.

Reason C5 refers to instances where the *cause* was not specified in the sentence but the computer program nevertheless extracted one part of the sentence as the cause. In some cases, the computer program was confused by the complex sentence structure. These errors can be avoided if an accurate parser is used. For some of the sentences, it is difficult to tell from the sentence structure alone whether the cause is specified or not. The following pairs of sentences illustrate this difficulty. The sentences labeled (a) do not specify the *cause*, whereas the sentences labeled (b) having the same syntactic structure do specify the *cause*:

(28a) Friends have suggested pouring [$_{\text{effect}}$ vermouth into the soap dispenser.]
(28b) [$_{\text{cause}}$ Friends] have admitted pouring [$_{\text{effect}}$ vermouth into the soap dispenser.]

(29a) Measures are being taken to make [$_{\text{effect}}$ the loan more attractive.]
(29b) [$_{\text{cause}}$ Low interest rates] are being offered to make [$_{\text{effect}}$ the loan more attractive.]

The most common reason for the errors of commission was word sense ambiguity. A word can have several senses, some senses having a causal meaning and others not. (No lexical disambiguation was attempted in this study.) Difficulty with complex syntactic structures (reasons C4 and C5) accounted for the next highest number of errors.

I now discuss the errors of omission. Reasons D1 to D3 (4) together accounted for the highest number of errors. These three types of errors can be reduced by using an accurate parser.

Reason D4 refers to sentences like the following:

(30) The flaps on each wing help *provide lift* for a jetliner to get off the ground.

In this sentence, the causal verb *lift* is used in a noun form. The sentence may be paraphrased as

(31) The flaps on each wing help *lift* a jetliner off the ground.

Nominalized verbs, i.e. verbs used in noun form, were not handled in this study.

I turn now to resultative constructions. As described in Chapter 3 (Section 3.4), I used the pattern

NP V NP Adj

to identify resultative sentences in which the resultative phrase is an adjective. This pattern picked out 8 causal relations in the sample sentences, 2 of which were correct and 6 incorrect. Of the 6 that were incorrect, 2 were due to errors in part-of-speech tagging. The two resultative sentences that were correctly identified are:

(32) For years, [$_{\text{cause}}$ Soviet officials] kept [$_{\text{effect}}$ the damning health statistics secret] while ...

(33) [$_{\text{cause}}$ Other provisions] would make [$_{\text{effect}}$ it more difficult for railroads to deny shippers the right to route their freight to competing railroads.]

Two of the sentences that were incorrectly identified as resultatives are:

(34) ... you can hold the lawyers responsible ...
(35) I'd call this plan mediocre at best.

From the high percentage of errors, I cannot recommend using the *NP-V-NP-Adj* pattern to identify resultative sentences. On examining the sentences picked up by the pattern, I found that all the verbs involved in these sentence, except one, have the grammar code *[+obj+adj]* in the *Longman Dictionary*. This suggests that most of the cases where an adjective can follow the object noun phrase is specified in the *Longman Dictionary*. We can therefore scan all the verb entries in the *Longman Dictionary* having a *[+obj+adj]* code, and make up a list of those verbs for which the adjective following the object noun phrase is likely to be a resultative phrase (call it List A) and a list of verbs for which the adjective is not a resultative phrase (call it List B). We can then either use List A to identify instances of adjective resultative phrases in text, or use List B to eliminate the instances where the adjective following the object noun phrase is not a resultative phrase. In the latter case, all other instances where the adjective appears after the object noun phrase can be identified as a resultative phrase.

What about other kinds of resultative phrases (e.g. resultative phrases in the form of a prepositional phrase)? Only one such resultative sentence was missed by the computer program. It appears that resultative constructions involving non-causal verbs are relatively rare in Wall Street Journal text.

4.7. Problems encountered by the judges in identifying causal relations in text

To gain some insight into the problems the judges had in identifying causal relations, each judge was asked to note the instances in the training sample where she had problem deciding whether there was a causal relation. These instances were later discussed with the judge. In addition, the causal relations identified by each judge in the training sample were compared with the causal relations that I had identified. Instances where the judge's decision differed from mine were discussed with the judge. To find out how the judge arrived at a particular decision, I asked the judge to think out loud why each sentence did or did not contain a causal relation.

The judges encountered three main problems:

1. It is sometimes difficult to decide whether a causal relation is explicitly expressed or merely implied.
2. It is not always obvious whether a verb does specify an effect or merely describes an action.
3. Because the sentences are presented for judgment out of context, it is sometimes not clear what the sentence means.

The distinction between explicitly expressed and implied causal relation is not clear cut. There are different degrees of explicitness. Also, because of lexical ambiguity, some inference from general knowledge is often necessary to identify the meaning or sense of a word. Sometimes, it is a matter of opinion whether a word/phrase is a causal link that explicitly indicates a causal relation, or is simply a conjunction with no causal meaning. Consider the following examples:

(36) [$_C$ The credit union lobby was very powerful in this state,] *and* [$_E$ the bills never got anywhere.]

(37) [$_E$ About half the bonds are held by Goldman, Sachs & Co.'s Water Street Corporate Recovery Fund, which acquired many of them] *after* [$_C$ a third toy maker, Mattel Inc., approached Goldman about making an offer for Tonka.]

The words *and* and *after* are usually not considered to be causal links. I would consider the causal relation in the above two examples to be implicit rather than explicit. However, one judge, with some justification, claimed that the words *and* and *after* have a causal meaning in the above sentences. The judges probably interpreted the instruction to identify "only the cause and effect that is *explicitly expressed*" as meaning "identify the cause and effect that is *clearly expressed*."

The judges sometimes had difficulty deciding whether the meaning of a verb included a specification of the effect or whether the verb merely described an action. Consider the following examples:

(38) Instituto Bancario San Paolo di Tonino, Italy's largest bank, *reported* 1990 net profit rose 8% at its parent bank to 605 billion lire from 560 billion lire the year before.

(39) IBM will certainly try to *trim* its prices ...

(40) Deutsche Aerospace said it hopes to *complete* the plan within the next few weeks.

In example (38), it is not clear whether the verb *report* denotes only the action of publishing some information or whether it also includes the meaning

cause to be known by the public.

In example (39), the phrase "trim its prices" can be interpreted as

the action of charging less for goods/services

or it can be interpreted as

cause the price of goods/services to be lower.

In example (40), the phrase "complete the plan" may be interpreted as

finish planning

indicating the end of the act of planning, or it can be interpreted as

cause a plan to exist.

Whether a verb is interpreted as causal or not may depend on the context. One judge said that she accepted the verb *report* as causal only when the reporter was "reporting" unusual events. Presumably, a reporter gets more credit for making known an event that most people don't know about, especially if the event is surprising. Causal attribution is, after all, the act of assigning credit or blame for some happening.

Here are three other examples which are amenable to both a causal and non-causal reading:

(41) Caloric's loss more than *offset* profitable performances at its Amana and Speed Queen divisions.

(42) The Swiss investment holding company's 1990 consolidated profit surged to 197.9 million francs from 4.6 million francs a year earlier, *when* results included a 200.6 million-franc write off of its indirect stake in the U.S. investment bank Drexel Burnham Lambert.

(43) But *so long as* this bill encourages quotas, and it does, *it should* be unacceptable no matter what comprise is offered.

Sentence (41) can be interpreted in either of the following ways:

1. The amount of Caloric's loss *is greater than* the amount of the profit from Amana and Speed Queen.
2. Caloric's loss *caused* the profit accruing from Amana and Speed Queen *to disappear*.

The first interpretation is "non-causal" whereas the second interpretation is a causal reading. Sentence (42) can have two possible readings. In one reading, the "200.6 million-franc write off" isn't really the cause of the surge in profit for the Swiss company. Including the 200.6 million-franc write off when calculating profit is merely a matter of accounting, i.e. juggling the figures. In another reading, the write-off *caused* the profit surge *to show up* in the accounts. Sentence (43) can have a "prescriptive" reading or a causal reading. In the prescriptive reading, the sentence prescribes that the bill, which encourages quotas, not be accepted. In the causal reading, the sentence is taken to mean: the fact that the bill encourages quotas should *cause people to reject it*.

Another problem mentioned by the judges was that the sentences were presented for judgment without the context of the article from which the sentences were extracted. This made it more difficult to understand the meaning of some of the sentences. Having read the training set of sentences and identified causal relations in them, I don't think this had a marked effect on the judgments. Even when we don't fully understand a sentence, we are usually able to say with some confidence whether there is a causal relation. The words and sentence constructions that indicate cause and effect do not seem to depend much on extra-sentential context for the reader to interpret their meaning. Furthermore, the computer program works under the same handicap -- the computer program does not make use of context in identifying causal relations in sentences.

4.8. Conclusion

The research question that this chapter seeks to address is:

Research question 1
How effectively can cause-effect information expressed in sentences be identified and extracted using linguistic patterns?

The results indicate that for Wall Street Journal text, about 68% of the causal relations that are clearly expressed within a sentence or between adjacent sentences can be correctly identified and extracted using the linguistic patterns developed in this study. Of the instances that the computer program identifies as causal relations, about 64% are correct (72%, if I give the computer the benefit of the doubt when the causal relation is not clearly wrong).

In this study, the causal relations that are clearly expressed in the text are taken as those identified by two human judges. The judges were asked to identify only the causal relations that were explicitly expressed. From their judgments, it is apparent that the judges interpreted "explicitly expressed" as meaning "clearly expressed." Their judgments included some causal relations that were clearly expressed but had to be inferred using general knowledge.

Most of the errors made by the computer program are due to

- complex sentence structure
- lexical ambiguity
- absence of inferencing from world knowledge.

This study makes use of a phrase bracketer but not a full parser. If an accurate parser is used, the maximum recall that can be attained is around 83% (assuming no error due to sentence structure), and the maximum precision attainable is about 82%. Much of the complexity of the linguistic patterns constructed in this study is due to the need to handle different sentence structures. If a parser is used, the linguistic patterns can be made much simpler, and fewer patterns need be used.

Accurate word sense disambiguation, especially for verbs, can also substantially reduce errors. Inferencing from world knowledge will also help, but it is possible to implement this only for very narrow domains.

How well will the approach used in this study work for other corpora? It depends, of course, on the corpus. Using linguistic patterns for identifying causal relations will be effective to the extent that the corpus satisfies the following conditions:

1. Most of the causal relations in the text are explicitly expressed using linguistic means. The reader is seldom required to infer cause-effect from general knowledge or domain knowledge.
2. Most of the sentences are simple and straightforward.
3. The subject content of the corpus is limited to a narrow subject area so that word sense ambiguity is not a problem (i.e. most words are used in only one sense in the corpus).

I surmise that the approach will work well with databases containing abstracts of journal articles in a particular subject area -- particularly abstracts reporting results of empirical research. Causal relations will probably be explicitly stated in such abstracts. I expect the approach to fare poorly with episodic text -- text describing a series of related events (e.g. a story). For this kind of text, causal relations between events usually have to be inferred by the reader using extensive knowledge about the types of events described in the text (Cullingford, 1978; Schank, 1982; Schank & Abelson, 1977; Wilensky, 1978).

CHAPTER 5
RETRIEVAL EXPERIMENTS

5.1. Introduction

This chapter describes the retrieval experiments that were carried out to investigate the second research question:

> **Research question 2**:
> Can the information obtained by matching causal relations expressed in documents with causal relations expressed in the user's query statement be used to improve retrieval effectiveness over just matching terms without relations?

Experimental results are reported and several subsidiary research questions are investigated.

The experimental retrieval system used in this study was based on the vector space model (Salton, 1989; Salton & McGill, 1983). The test collection used was a subset of the TIPSTER/TREC test collection (Harman, 1993b & 1993c). Only the *Wall Street Journal* documents that had relevance judgments from TREC-1 and TREC-2 conferences were used in this study. Retrieval effectiveness was measured using the normalized recall, normalized precision, the 11-point recall-precision average, and the 3-point recall-precision average.

Each retrieval experiment had the following two steps:

1. *a model building or exploratory step*, carried out to determine the best set of weights to use for combining the scores from the different types of matching,
2. *a model validation or confirmatory step*, carried out to determine the retrieval effectiveness of the models developed in the model building step.

Two types of retrieval experiments were carried out:

1. ad hoc queries experiment
2. routing queries experiment.

The *ad hoc queries* experiment simulated the retrieval situation where a predetermined retrieval strategy is applied to all queries, and no relevance feedback information is used to reformulate the query or modify the retrieval strategy. The model building step of this experiment was carried out using 39 query statements, and the model validation step was performed using a different set of 38 query statements.

The *routing queries* experiment simulated the situation where a search profile is developed for an SDI (Selective Dissemination of Information) service. In this situation, a search is carried out on the available database and an SDI search profile is developed based on the user's relevance judgments on the retrieved documents. Subsequently, all new documents entering the database are matched against the SDI profile to determine if the documents are likely to be relevant to the user. For this routing queries experiment, the best set of weights to use was determined for each query using one half of the database and then validated using the other half of the database. The results obtained in this experiment thus represented the best possible results obtainable in the ad hoc queries experiment using the same retrieval method. The retrieval improvement obtained in the ad hoc queries experiment cannot, in theory, exceed the improvement obtained in the routing queries experiment, since in the latter the weights used were optimized for individual queries.

5.2. Research Questions

5.2.1. Main research question

The research question that is addressed in this chapter is:

> **Research question 2**:
> Can the information obtained by matching causal relations expressed in documents with causal relations expressed in the user's query statement be used to improve retrieval effectiveness over just matching terms without relations?

To investigate this question, the linguistic patterns described in Chapter 4 were used to identify causal relations expressed in documents. The terms and causal relations identified in the documents were then matched with terms and causal relations expressed in users' query statements. If a query states that the desired documents should express that *A* causes *B*, then documents that indicate that *A* causes *B* should receive higher retrieval scores, reflecting a closer match with the query, than documents mentioning *A* and *B* but without the causal relation linking them.[25] The research question is whether scoring the match between the query statement and the documents using information from causal relation matching (in addition to information from term matching) will yield better retrieval results. Another way of looking at this is: does the degree of causal relation match between a query and a document capture some similarity between the query and the document not captured by term matching?

Several subsidiary questions that are related to *research question 2* are investigated in this chapter. The subsidiary questions relate to

- the use of Roget category codes for information retrieval
- the characteristics of the documents
- the characteristics of the queries
- the use of term proximity as a substitute for identifying relations in documents.

The subsidiary questions are discussed in the next sections.

5.2.2. Use of Roget codes

The use of Roget category codes for recall enhancement is an interesting feature of this study. Substituting Roget codes for words in documents and queries results in a greater number of term matches between documents and queries. It will allow a word in a query to match with synonymous words in the document. It will also allow variant forms of the same word to match. However, a word can have more than one meaning, each meaning having its particular Roget code. Since no lexical disambiguation is attempted in this study, the use of Roget codes will result in a greater number of *incorrect* term matches. The question is whether the advantage of having a greater number of *correct* term matches outweighs the disadvantage of also having a greater number of *incorrect* matches:

> **Subsidiary question 2.1a**
> Does the use of Roget codes *in place of keywords* improve retrieval effectiveness?

[25] It is assumed that during retrieval, the system calculates a score for each retrieved document indicating the degree of match between the document and the query statement. A higher score indicates a closer match. Documents can then be ranked in decreasing order of the score.

Subsidiary question 2.1b
Does the use of Roget codes *in addition to keywords* improve retrieval effectiveness over using keywords alone?

Subsidiary question 2.1c
Does the use of Roget codes *in addition to keywords* as the terms in causal relations improve retrieval effectiveness?

5.2.3. Characteristics of the document

Causal relation matching may be more effective for some documents than others in determining the similarity between the document and the query. If we can identify the documents for which causal relation matching is likely to help in predicting relevance and the documents for which causal relation matching is not likely to help, then we can weight causal relation matches differently for different documents and this will give better results than using the same set of weights for all the documents.

One possible factor is the degree of term match (i.e. keyword and/or Roget code match) between the document and the query. If the document already has a high score from term matching, causal relation matching might not tell us much more about the similarity between document and query. If this is true, then the higher the score from term matching, the smaller the weight that should be given to the score from causal relation matching in calculating the composite retrieval score.

Subsidiary question 2.2:
Is there an interaction between term matching score and causal relation matching score in determining the relevance of a document?

5.2.4. Characteristics of the query

The effect of causal relation matching may also depend on some characteristics of the query. One possible factor is whether the causal relation is an important part of the query or of secondary importance in the query. The improvement in retrieval result from causal relation matching may be greater when the causal relation is central to the query than when the causal relation is peripheral to the query (i.e. not an important part of the query).

Subsidiary question 2.3a:
Is the improvement in retrieval effectiveness from causal relation matching greater when the causal relation is *central* to the user's query than when the causal relation is *peripheral* to the query?

Another factor that probably affects the efficacy of causal relation matching is what I shall refer to as the "strength" of the causal association. There are some pairs of terms that if they co-occur in a document we can safely assume there is some kind of causal relation between them and we can confidently predict in which direction the causal relation goes. Take the following pairs of terms:

smoking	cancer
drug	AIDS

If the terms "smoking" and "cancer" both occur in a newspaper article, it is very likely that the relation "smoking causes/caused cancer" is expressed or implied in the article.[26] If the terms "drug"

[26] The relation may, of course, be negated, as in this example:

and "AIDS" occur in the same article, it is very likely that the "drug" referred to is supposed to have some causal effect on the condition of the AIDS patient. In such cases, it is not necessary to perform complex natural language processing to find out whether the causal relation is expressed in the document. The presence of the two terms already implies the causal relation, and term matching should be sufficient for determining similarity between document and query. Causal relation matching will not provide any additional information in this case.

> **Subsidiary question 2.3b**:
> Is the improvement in retrieval effectiveness from causal relation matching greater when the causal association between the two terms in the relation is weak than when the association is strong?

How can the strength of the causal association between two terms be determined? One method is to examine a sample of documents containing the two terms and then find out the percentage of documents in which the causal relation is expressed or implied. This procedure is time consuming. Another approach is to make use of people's subjective impressions of the strength of the association. There are several ways of eliciting people's subjective impressions. In this study, I simply used my own subjective impression of the strength of the association. For each causal relation, I wrote down a number between 0 and 10 to express my subjective estimate of how likely the causal relation was expressed or implied in a document given that most of the keywords in the causal relation occurred in the document. Since I did the subjective estimation myself, the results obtained are very preliminary.

An indirect way of exploring this issue is based on the idea that the strength of the causal relation will be partially reflected in the retrieval result from term matching. The retrieval result from term matching should be better when the causal association is strong than when the causal association is weak. Causal relation matching should have a greater impact on queries that do poorly with term matching.

> **Subsidiary question 2.3c**:
> Is the improvement in retrieval results from causal relation matching greater for queries that obtain poor retrieval results from term matching than for queries that obtain good results from term matching?

If a query does not show an improvement in retrieval result from causal relation matching, it may be because the automatic method used for identifying causal relations is not effective for identifying the causal relation specified in the query. Cause-effect can be expressed in many different ways, and a particular causal relation may tend to be expressed in a certain way. If a causal relation tends to be expressed in a way that is not effectively handled by the computer program developed in this study, then causal relation matching will not improve the retrieval result for the query. This study focused on causal relations expressed within the sentence or between adjacent sentences. Not handled were causal relations occurring across larger distances -- between different paragraphs or between different sections of a document. Also not handled were causal relations that were not explicitly expressed in a document, but were meant to be inferred by the reader using general knowledge. So, if causal relation matching didn't improve the retrieval results for a particular query, it might be because the automatic method for identifying causal relations in this study was not good at identifying the type of cause-effect expressed in the query.

The difficulty of identifying the causal relation of interest would be partially reflected in the total number of causal relation matches found when matching the query with documents in the database. In other words, when matching a query with documents in a database, we may be able to

---

Cigarette manufacturers deny that smoking causes cancer.

predict whether causal relation matching will improve retrieval results by the number of causal relation matches found.

Subsidiary question 2.3d:
Is the improvement in retrieval results from causal relation matching greater for queries for which there are more causal relation matches found during retrieval than for queries with fewer causal relation matches?

To answer this question, two measures were used to measure the amount of causal relation matches:

- sum of the causal relation match scores for all the documents
  total number of documents in the database

- sum of the causal relation match scores for all the documents
  total number of *relevant* documents in the database

5.2.5. Use of term proximity

Identifying causal relations in text is computationally expensive. If causal relation matching is found to be effective in improving retrieval effectiveness, it is important to find out whether the simpler method of using *term co-occurrence within a sentence* (i.e. term proximity) will give equally effective results. In other words, if a user wants documents that say that A causes B, can we safely assume that the causal relation between A and B is expressed in the document simply from the fact that term A and term B occur within the same sentence in the document?

Subsidiary question 2.4a:
Is the improvement in retrieval results obtained using causal relation matching greater than the improvement obtained using term proximity matching (i.e. allowing causally related terms in the query statement to match with terms that co-occur within document sentences)?

If term proximity is found to give the same or better results than causal relation matching, perhaps using both causal relation matching *and* term proximity matching together will give better result than using either. In other words, causal relation matching may capture some similarity between query and document not captured by term proximity matching. Using term proximity, the retrieval program will be able to pick out all the causal relations, including those missed by the automatic method used in this study for identifying causal relations. However, term proximity matching may make more errors in picking out co-occurring terms that are not causally related. Also, this study makes use of causal relations of the form word -> *

* -> word

where "*" is a wildcard that can match with anything. This is described in detail later. This type of causal relation matching where only one member of the relation (either the cause or the effect) need to match is not approximated by term proximity matching where both terms are required to occur in a sentence.

Subsidiary question 2.4b:
Does the use of causal relation matching *in addition* to term proximity matching yield better retrieval results than using term proximity matching alone?

5.3. The Test Collection, Retrieval System, and Evaluation Method

<top>

<head> Tipster Topic Description

<num> Number: 081

<dom> Domain: Finance

<title> Topic:

Financial crunch for televangelists in the wake of the PTL scandal

<desc> Description:

Document will report a loss of revenue of a televangelist in the aftermath of the PTL scandal, or a financial crisis triggered by the scandal.

<narr> Narrative:

A relevant document will identify a religious broadcaster, and report a specific dollar amount or a percentage loss, indicating a decline in revenue suffered by that broadcaster as a result of consumer reaction to the Jim Bakker scandal.

<con> Concept(s):

1. donations, confidence, checks, money, dollars

2. plead, beg, cry

3. down, fall, slip

4. religious broadcasters, televangelists

5. Bakker, Falwell, Robertson, Swaggart, Schuller, Roberts

<fac> Factor(s):

<def> Definition(s):

</top>

Figure 3. Example of a TIPSTER/TREC query statement

5.3.1. The test collection

The test collection used in this study was a subset of the test collection used in the ARPA[27] TIPSTER project (Harman, 1993b) and used by the participants of TREC-1 and TREC-2 conferences (Harman, 1993c; Harman, 1994b).[28] The TIPSTER/TREC document collection includes

[27] Formerly known as DARPA (Defense Advanced Research Projects Agency).

[28] *TREC* is the acronym for *Text REtrieval Conference*. There has since been a third TREC conference

records from the following sources: Wall Street Journal, AP Newswire, Computer Select (Ziff-Davis Publishing), Federal Register, Abstracts from the Department of Energy (DOE), San Jose Mercury News and U.S. Patents.

The test collection includes 150 query statements.[29] The short titles of these queries are listed in Appendix 5. An example of a query statement is given in 3. The queries were constructed by actual users of a retrieval system in one of the government agencies. As can be seen in Appendix 5, the queries are on topics that are generally covered by newspapers. Most of the queries are on business and politics-related topics, international affairs, health issues and science topics of general interest. The query statements are relatively long, compared to those in other test collections, and typically contain a few concepts and relations (i.e. they are not one-concept queries). The query statements are fairly detailed specifications of what information a document must contain in order to be relevent.

Relevance judgments for these 150 queries were obtained by the organizers of TREC-1 and TREC-2 conferences for the top ranked documents retrieved by each participating information retrieval system. 25 retrieval systems participated in TREC-1 and 31 systems in TREC-2. Some of the participating retrieval systems submitted more than one set of results, each using a different retrieval strategy. For each query statement, the top ranked 100 documents from each set of results submitted by participating systems were pooled, and TREC organizers obtained relevance judgments for these documents.

The relevance judgments were not performed by the same people who constructed the queries. However, all the relevance judgments for a query were done by the same judge. The judges were not given any general instructions on how to assess relevance. The query statements themselves were used as the instructions to judges. The relevance judgments were probably made on the basis of "conceptual relatedness" rather than "usefulness," since the query statements do not state how the information obtained would be used. A document was judged relevant if it contained the desired information somewhere in the document -- even if the information was found in only one short paragraph in a long document. In other words, relevance was assessed not in terms of the general topic of the document but on whether it contained the information specified in the query statement.

This study made use only of the set of Wall Street Journal articles (1986-1992) that had relevance judgments from TREC-1 and TREC-2 conferences. The number of Wall Street Journal articles that had relevance judgments for each query is indicated in Appendix 5. For each query, a test database was constructed comprising the Wall Street Journal articles (full-text) that had relevance judgments *for the query*. In other words, a different test database was used for each query. Since the test database for each query consisted of the top 100 documents retrieved by other systems, the experiments were in fact testing whether causal relation matching could be used as a precision-enhancing procedure to make fine distinctions among the top-ranked documents from other retrieval systems.[30]

Wall Street Journal documents were selected for this study because I had worked with

(Harman, 1995).

[29] An additional 50 query statements were added to the test collection for the TREC-3 conference (Harman, 1985) and another 50 for TREC-4.

[30] Pooling the top ranked documents from several retrieval systems can be seen as a recall-enhancement strategy for identifying most of the documents that are likely to be relevant. Most of the systems that participated in TREC-1 and TREC-2 used some variation of keyword matching (usually with query expansion). Very few used relation matching.

this collection in the DR-LINK project at Syracuse University (Liddy & Myaeng, 1993; Liddy & Myaeng, 1994; Myaeng & Liddy, 1993; Myaeng, Khoo, & Li, 1994), and so was familiar with it. Also, a considerable amount of text processing was carried out on the documents in the DR-LINK project, and I made use of the processed text in this study. Many types of text processing were done in the DR-LINK project, but the processing steps relevant to this study are as follows:

- sentence boundaries were identified and each sentence stored on a separate line
- words were tagged with part-of-speech tags (using the POST tagger (Meteer, Schwartz, & Weischedel, 1991) obtained from BBN Systems and Technologies)
- phrases and clauses were bracketed and labeled.

Sample processed text is given in Appendix 2. In this study, pairs of phrases that were causally related were extracted from the documents using the computer program and linguistic patterns developed in this study and described in Chapter 4.

Of the 150 query statements, 78 contain one or more causal relations, and these were the queries used in this study. The 78 queries containing causal relations are indicated in Appendix 5. Causal relations in the query statements were manually identified by me. The linguistic patterns developed in this study for identifying causal relations in Wall Street Journal text could not be applied to the query statements for the following reasons:

- natural language query statements have their own particular sublanguage. The linguistic patterns have to be specially tailored for query statements.
- some query statements have causal relations in which the cause or the effect is not specified and is supposed to be supplied by the relevant document. For example, in the query

  I'm interested in documents that describe the consequences of the Gulf War.

  only the cause *Gulf War* is specified. The effect of the Gulf War is not stated in the query but must be stated in the document for the document to be considered relevant. The causal relation in the query can be represented as

  Gulf War -> *

  where "*" is a wildcard that can match with anything. The wildcard is like a question mark or a blank to be filled in. The linguistic patterns developed in this study were constructed to identify causal relations in which both the cause and effect are specified.

To ensure that I did not overlook any causal relation in the query statements, two other sets of "judgments" were obtained. One set of judgments was obtained from the computer program developed in this study for identifying causal relations. Another set of judgments was obtained from one of the two human judges (judge A) who provided judgments for evaluating the effectiveness of the computer program (as reported in Chapter 4). Each causal relation identified by the human judge or computer program but not by me was reviewed and added to the list of causal relations if it appeared to me to be reasonable.

For each causal relation found in a query statement, I also decided whether the causal relation was *central* to the query (i.e. was an important part of the query) or *peripheral* to the query (i.e. of secondary importance). 52 of the queries contain a causal relation which I considered to be central to the query. These are indicated in Appendix 5.

I stated earlier that two types of retrieval experiments were carried out in this study:

1. ad hoc queries experiment
2. routing queries (or SDI) experiment.

Each experiment was divided into two stages:

1. *a model building or exploratory step*, which was carried out to determine the best set of weights to use for combining the scores from the different types of matching.
2. *a model validation or confirmatory step*, which was performed to find out the retrieval effectiveness of the models developed in step 1.

In the *ad hoc queries* experiment, 39 queries were used for model building, and a different 38 queries were used for model validation.[31] The queries used for model building were selected simply by taking every other query from the set of queries in which the causal relation was central to the query, and every other query from the set in which the causal relation was peripheral. The queries used for model building and model validation are indicated in Appendix 5.

For the *routing queries* (or SDI) experiment, I made use of 72 queries.[32] For each query, half of its test database (articles from 1986, 1987 and 1991 Wall Street Journal) was used for model building. The other half of the test database (articles from 1988, 1989, 1990 and 1992) was used for model validation. My purpose in carrying out the routing queries experiment was to see how much retrieval improvement could be achieved for individual queries using causal relation matching if the weights used for combining scores were optimized for each query. It is stressed that, unlike what is usually done in routing queries experiments, the content of the *relevant* documents was not used to reformulate the query. Only the weights used for combining the scores from the different types of matching were optimized.

5.3.2. The retrieval system

5.3.2.1. Overview

The retrieval system used in this study was based on the *vector space model* (Salton, 1989; Salton & McGill, 1983). This model is often used in information retrieval research and is well-known to give relatively good retrieval performance (Salton, 1989). In vector-based retrieval systems, each query and each document is represented by a vector, which is composed of a sequence of weights, each weight representing the quantity of some feature of the query or document. The degree of match between a query and a document is taken to be the similarity between the query and the document vector -- the similarity being calculated using one of several similarity measures.

The experimental retrieval system used was developed as part of this study and was written in the C programming language. I wrote the programs myself, except for the stemming function (which was written by a programmer in the DR-LINK project) and the evaluation program for generating recall-precision figures (which was obtained from the SMART system (Salton & McGill, 1983)). This section describes how the vector space approach to information retrieval was implemented in my experimental retrieval system.

The query and document vectors in this study were determined by the following three factors:

- *type of matching*. Eight types of matching were used: term matching, term proximity matching, and six types of causal relation matching

[31] One of the 78 queries containing causal relations was dropped from the experiment because no document from Wall Street Journal was found to be relevant to the query in TREC-1 and TREC-2 conferences.

[32] Five of the 78 queries containing causal relations were dropped from this experiment because there was no *relevant* document in the half of the database used for model building or in the half of the database used for model validation. One query was dropped because of an error in a processing step.

- *type of term conflation*. Two types of term conflation were used: converting each word to its base form, and replacing each word with its Roget category codes
- *weighting scheme* for calculating the component weights in a vector.

For each query and each document, the retrieval system constructed a separate sub-vector for each combination of the above three factors -- type of matching, type of term conflation, and weighting scheme.

When matching a query with a document, each query subvector was matched with a document subvector, and a similarity score was calculated. The system then combined all these subvector similarity scores into a composite similarity score by taking a weighted sum of the subvector similarity scores. As previously mentioned, the best set of weights to use when combining the subvector similarity scores was determined empirically in the model building step of the experiments, using half of the queries in the case of the ad hoc queries experiment and using half of the test database in the case of the routing queries experiment.

The approach of using several subvectors to represent documents and queries is similar to the one used by Fagan (1989). As Fagan pointed out, combining the *term subvector* and *relation subvector* into one long vector and calculating a similarity score between the extended query and document vectors can result in the anomalous situation where relation matches actually reduce the similarity score.[33] By separating the *term* and *relation subvectors*, we can ensure that relation matches can only increase the similarity score.

The following sections describe in greater detail each aspect of the retrieval process.

5.3.2.2. Types of matching

The types of matching performed in this study were as follows:

- *term matching*, in which terms in the document were matched with terms in the query.
- *causal relation matching*, in which pairs of causally related terms in the document were matched with pairs of causally related terms in the query statement. Six types of causal relation matching were carried out.
- *term proximity matching*, in which pairs of terms that co-occurred within a sentence in the document were matched with pairs of causally related terms in the query statement.

The different types of matching were implemented by constructing the query and document subvectors in appropriate ways, as described below.

Term matching was implemented by having each component of a query/document subvector correspond to each unique term in the document collection. Each component weight of this *term* subvector represented the quantity of a particular term in the query or document. The component weight was calculated using a weighting scheme. One possible weighting scheme is to simply take the number of times the term occurs in the query or document as the component weight. The weighting schemes used in this study are discussed later.

Causal relation matching matched the pairs of causally-related phrases in the query statement with pairs of causally related phrases that had been identified in the document. Causal relation matching was done in six ways:

[33] This assumes that cosine normalization is applied to the query and document vectors, or that the cosine similarity measure is used.

1. "term -> term" matching
2. "termpair -> term" and "term -> termpair" matching
3. "term -> *" matching
4. "* -> term" matching
5. "termpair -> *" matching
6. "* -> termpair" matching

A separate subvector was constructed for each of the six types of causal relation matching.

"Term -> term" matching entailed generating from each pair of causally related *phrases* every possible pair of *terms*, one term from the *cause* phrase and the other term from the *effect* phrase. As an example, suppose the following causal relation was found in the query statement:

cigarette smoking -> lung cancer

Then, the following "term -> term" pairs would be generated:

cigarette -> lung
cigarette -> cancer
smoking -> lung
smoking -> cancer

The "term -> term" pairs can be viewed as automatically generated index phrases. Each of these "term -> term" pairs was represented by a component in the "term -> term" subvector constructed for each query and document. Each component weight of a "term -> term" subvector represented some quantity of a particular "term -> term" pair in the query or document (e.g., the number of times the "term -> term" pair occurred in the query statement or document).

For "termpair -> term" matching, each term in the *cause* phrase was paired with each of the other terms in the *cause* phrase to form *termpairs*, and each of these termpairs was linked to each term in the *effect* phrase. In other words, each "termpair -> term" triple was constructed by taking two terms from the *cause* phrase and one term from the *effect* phrase. As an example, from the causal relation

cigarette smoking -> lung cancer

the following "termpair -> term" triples would be generated:

cigarette smoking -> lung
cigarette smoking -> cancer

"Term -> termpair" triples were generated in a similar way. One term was taken from the *cause* phrase and two terms were taken from the *effect* phrase:

cigarette -> cancer lung [34]
smoking -> cancer lung

Each "termpair -> term" and "term -> termpair" triple was represented by a component in a query/document subvector. "Termpair -> term" and "term -> termpair" matching, since they involved

[34] Word order in a term pair is not significant. In the study, the terms within a term pair are sorted alphabetically.

three terms from each causal relation, provided more precise matching than "single term -> single term" matching.

For "term -> *" matching, terms were extracted from the *cause* phrase only. From the example relation "cigarette smoking -> lung cancer", the following "index phrases" would be generated:

cigarette -> *
smoking -> *

"*" is a wildcard symbol indicating that the effect term is unspecified. Thus each term that occurred in a *cause* phrase was represented by a component in the "term -> *" subvectors. "Term -> *" matching is less constrained than "term -> term" matching. In effect, terms occurring in a *cause* phrase in the query statement were matched with terms occurring in the *cause* phrases in the document. This kind of matching would allow the relation

cigarette smoking -> lung cancer

in a query statement to match

cigarette smoking -> respiratory disorders

in a document.

For "* -> term" matching, terms were extracted from the *effect* phrase instead of the *cause* phrase. From the example "cigarette smoking -> lung cancer" the following "index phrases" would be generated:

* -> lung
* -> cancer

Thus each term that occurred in an *effect* phrase was represented by a component in the "* -> term" subvectors. In effect, terms occurring in an *effect* phrase in the query statement were matched with terms occurring in the *effect* phrases in the document. This kind of matching would allow the relation

cigarette smoking -> lung cancer

in the query statement to match

air pollution -> lung cancer

in a document.

For "termpair -> *" matching, pairs of terms were extracted from the *cause* phrase. From the example "cigarette smoking -> lung cancer", the following "index phrase" would be generated:

cigarette smoking -> *

Similarly, for "* -> termpair" matching, pairs of terms were extracted from the *effect* phrase. From the example "cigarette smoking -> lung cancer", the following "index phrase" would be generated:

* -> cancer lung

I have described six ways in which causal relation matching was implemented in this study. "Term -> *" and "* -> term" matching is the least constrained. If a term in a *cause* phrase in the query also occurred in a *cause* phrase in the document, it was considered a partial match. Similarly, there was a partial match if a term in the *effect* phrase in the query also occurred in an *effect* phrase in the document. The other kinds of matching

termpair -> *
* -> termpair
term -> term
term -> termpair and termpair -> term

gave more precision since they required two or more terms from a causal relation in the query to co-occur in a causal relation in the document before it was considered a partial match. There are other possible ways of implementing relation matching within the vector space model. However, the approach described above is probably the simplest and easiest to implement.

The last type of matching used in this study was *term proximity matching*. For term proximity matching, pairs of terms were generated from the documents and the query statements. The pairs of terms were constructed differently for documents than for queries. For documents, each term in a sentence was paired with all the other terms in the same sentence, so the pairs were composed of terms that co-occur in a sentence. For query statements, the pairs of terms were constructed by pairing each term in a *cause* phrase with each term in the corresponding *effect* phrase. From the example "cigarette smoking -> lung cancer", the following pairs of terms would be generated (the members of each pair being arranged in alphabetic order):

cigarette - lung
cancer - cigarette
lung - smoking
cancer - smoking

Each unique pair of terms generated from the sentences in the document collection and from the cause-effect phrases in query statements was represented by a component in the *term proximity* subvectors. The effect of constructing the subvectors in this way was that pairs of causally related terms in a query statement were allowed to match pairs of terms that co-occurred within a document sentence. This type of matching assumes that when two terms are causally related in a query statement, then they are also causally related in a document if they co-occur within a sentence in the document.

5.3.2.3. Types of term conflation

The following two *methods of term conflation* were used in this study:

- converting each word to an entry word in the machine readable version of *Longman Dictionary of Contemporary English* (2nd ed.)
- replacing each word with category codes from *Roget's International Thesaurus* (3rd ed.).

Conversion of words to an entry word in *Longman Dictionary* was performed using a computer program developed in the DR-LINK project (Liddy & Myaeng, 1993). The program performs this function by identifying potential suffixes in a word and replacing the suffixes until the word matches an entry in *Longman Dictionary*. If the word is a verb, it is first checked against a list of irregular verb forms. If found, the irregular verb (e.g., *said*) is replaced by its corresponding base verb (e.g., *say*). This conversion program can be considered to be a "weak" stemmer. It conflates singular and plural forms, and word forms for the different tenses. However, it does not conflate words belonging to different grammatical categories if each has a separate entry in the dictionary. For example, the words "translate", "translation" and "translator" each has a separate entry in the *Longman Dictionary* and so are not conflated. Words in proper nouns are stemmed in the same way but are converted to uppercase so that when vectors are constructed, stemmed words from proper nouns are considered different terms from the same words that are not proper nouns.

The second method of term conflation used in this study was to replace words in the text with Roget codes. Only nouns, verbs, adjectives and adverbs were replaced with Roget codes and used as terms in the query/document subvectors. If a word did not have a code in *Roget's International Thesaurus*, then the stemmed word itself was treated as a Roget code when constructing the query/document vectors.

A word may have more than one Roget code -- a different Roget code for each sense of the word and each grammatical category the word belongs to (e.g., *cost* can be a verb or a noun). In this study, no word sense disambiguation was attempted and the Roget codes for all the senses of a word were used when constructing the vectors. However, since each word in the text had already been tagged with part-of-speech labels in an earlier text processing step, it was possible to select only those Roget codes that were applicable to the word's part-of-speech category.

The classification scheme in *Roget's International Thesaurus* has seven hierarchical levels: class, subclass, heading, category, paragraph, and semicolon group. When using Roget codes for term conflation, one can select which level of the Roget classification scheme to use. The *semicolon group* code is the most specific code. Higher level codes allow more words to be conflated and are likely to improve recall to the detriment of precision. It was decided to use the *category code* for this study because it conflates not only synonymous words belonging to the same grammatical category but also related words from different grammatical categories.

5.3.2.4. Weighting schemes

		weighting scheme for	weighting scheme for
code	type of matching	query vector components	document vector components
k_1	keyword	ntf*idf cos	tf*idf cos
k_2	keyword	bin cos	bin*idf
r_1	Roget code	ntf*idf cos	tf*idf cos
r_2	Roget code	bin cos	bin*idf
v	keyword->*	bin norm	bin*idf
vp	keyword pair->*	bin norm	bin*idf
vr	Roget code->*	bin norm	bin*idf
w	*->keyword	bin norm	bin*idf
wp	*->keyword pair	bin norm	bin*idf
wr	*->Roget code	bin norm	bin*idf
c	keyword->keyword	bin norm	bin*idf
cp	keyword pair->keyword & keyword->keyword pair	bin norm	bin*idf
cr	Roget code->Roget code	bin norm	bin*idf
a	keyword pair within sentence	bin norm	bin*idf
ap	keyword triple within sentence	bin norm	bin*idf
ar	Roget code pair within sentence	bin norm	bin*idf

Abbreviations

bin binary weighting: the value 1 is assigned as the component weight if the feature is present in the query or document, 0 otherwise.

tf term frequency weighting: the value assigned is the number of times the feature occurs in the query or document.

ntf normalized term frequency weighting: a variation of the term frequency weighting in which the weight is constrained to vary between the values 0.5 and 1.0. The weight is calculated using the formula: 0.5 + 0.5 * tf/(max tf)

idf inverse document frequency, calculated using the formula *LOG (N/df)* where *N* is the total number of documents in the document collection and *df* is the document frequency of the term (the number of documents containing the term).

cos cosine normalization: dividing each component weight by the length of the vector, which is given by the formula $\sqrt{[\ \Sigma^{n}_{i=1}\ \text{weight}_i^2\ ]}$, where *n* is the number of components in the vector and $weight_i$ is the weight for component *i* of the vector.

norm normalization by dividing each component weight by the length of the k1 query vector for relations between keywords and r1 query vector for relations between Roget codes.

Table 5. Similarity sub-scores generated for each document during retrieval

All the subvectors that were constructed for each query and document are summarized in 5. There was a subvector for each combination of 1) type of matching and 2) conflation method (i.e. word stemming or replacing each word with Roget codes). Each line in 5 describes how a query and a document subvector was constructed. Column 1 in the table gives a code for each query and document subvector. Column 2 indicates which type of matching and conflation method the subvector represents. Column 3 gives the weighting scheme used to calculate the component weights of the query subvector. Column 4 gives the weighting scheme used to calculate the component weights of the document subvector. So the weighting scheme used to calculate the vector component weights was different for query subvectors and document subvectors. When matching a query with a document, each query subvector was matched with the corresponding document subvector as listed in 5. I shall use the code given in the first column of the table to refer to the subvector similarity score generated by matching the pair of query and document subvectors. The subvector similarity scores thus generated were combined to obtain a composite similarity score for the document and query.

In this section, I discuss the weighting schemes used to determine the component weights of the subvectors. In the next section, I shall discuss the similarity measures used to calculate the subvector similarity scores, and how the subvector scores were combined to obtain the composite similarity score.

For term matching (i.e. keyword matching and Roget code matching), the weighting scheme used for the query subvector was *ntf*idf* with cosine normalization,[35] and the weighting scheme used for the document subvector was *tf*idf* with cosine normalization. In 5, term matching using these weighting schemes are labelled k_1 for keyword matching and r_1 for Roget code matching. 5 also gives the mathematical formulas for the weighting schemes. In a preliminary experiment, this pair of weighting schemes for the query and document vectors was found to give the best results for keyword matching among the well-known weighting schemes. This pair of weighting schemes was also found to give relatively good results for Roget code matching. The results of this preliminary experiment are given in Appendix 6.

The weighting schemes *ntf*idf* and *tf*idf* make use of the term frequency of each term -- the number of times the term occurs in the query or document. There are two ways of calculating the term frequency of a Roget code. The first approach takes into account the number of Roget codes assigned to a word. (Each word in Wall Street Journal is assigned an average of about five Roget codes.) If a word has *n* Roget codes, then this approach assigns the weight *1/n* to each of the codes assigned to the word. When counting the term frequency of a Roget code, this occurrence of the Roget code is counted as *1/n* instead of as *1*. The second approach of calculating the term frequency of a Roget code is to assume that all Roget codes assigned to words in the query or document have the same weight. Term frequency is obtained simply by counting the number of times the code occurs in the document or query. The retrieval results for these two ways of determining the term frequency of Roget codes is given in Appendix 6, Section B and C. Retrieval results were better for weighted Roget codes for all the schemes except the schemes that use *tf*idf* with cosine normalization for document terms and either *ntf*idf* or *tf*idf* for query terms (column 16 and 17 in Appendix 6). The retrieval results were nearly the same in these two cases. The combination of using *idf* (inverse document frequency) and cosine normalization appears to offset the disadvantage of using unweighted Roget codes. In this study, I used the first approach of counting weighted Roget codes. As I have just indicated, using weighted or unweighted Roget codes doesn't seem to matter for the weighting schemes that I used.

For the different types of causal relation matching as well as for the term proximity matching, the weighting scheme used for the query subvectors was binary weighting with normalization, and the weighting scheme used for the document subvectors was *binary*idf* weighting. The "normalization" carried out on the query subvectors is explained later. In a preliminary experiment, this pair of weighting schemes was found to be the most promising one among the

[35] The *ntf*idf* weighting scheme was recommended by Salton & Buckley (1988) for short query statements.

commonly used schemes.

Note that the weighting schemes used for term subvectors k_1 and r_1 were different from the weighting schemes used for the subvectors constructed for causal relation matching. The weighting schemes for the term subvectors included term frequency as a factor in the weighting schemes, whereas the weighting schemes used for the relation subvectors involved binary weighting. In this study, the subscores from term matching were combined with the subscores from relation matching. If combining the subscores from term and relation matching produced an improvement in retrieval results, part of the improvement might be due to the fact that different weighting schemes were used for term and relation matching. In other words, part of the improvement might be due to the fact that binary weighting captured some similarity between a document and query not captured by the term frequency weighting used for term matching.[36] It might be that were binary weighting also used for term matching, the improvement from causal relation matching would be less or none. To ensure that any improvement from causal relation matching could not be ascribed to the use of different weighting schemes, I introduced two other term subvectors (labeled k_2 and r_2 in 5) which used binary weighting. So, the baseline retrieval results in this study were the retrieval results for keyword matching using both the k_1 and k_2 subvectors. Retrieval improvements from causal relation matching were calculated in comparison with this baseline.

Normalization was performed on each query subvector. Usually, normalization of the query vector has no effect on the retrieval results. Normalization of the query vector affects all documents equally and so have no impact on the ranks of the documents. However, this is true only when a single vector is used to represent each query and each document. In this study, each query and document was represented by several subvectors. During retrieval, a subscore was obtained for each document subvector by matching the document subvector with a query subvector. A composite score was then computed for the document by taking the weighted sum of the subscores. The appropriate weight to use for each subscore was determined empirically. In this situation, whether the subvectors are normalized or not can affect the ranks of the documents. Normalization of the query subvectors can change the relative weight assigned to each subscore.[37]

For the term subvectors k_1, k_2, r_1 and r_2 for each query, cosine normalization was performed (by dividing the component weights of each subvector by the length of the subvector). The effect of normalization is that each component weight of a query subvector now reflects the importance of the term relative to the whole set of terms in the query (i.e. the length of the query vector). A term has a greater weight in determining relevance when there are few terms in the query statement than when there are many query terms.

For the *causal relation* subvectors for each query, I feel that it is not appropriate to use cosine normalization. This is because there are usually a few other relations in a query besides the causal relation. Cosine normalization of a *causal relation* vector will not tell us how important each causal relation is with respect to the whole set of relations in the query. In this study, I perform some kind of "normalization" by dividing the component weights of the *causal relation* subvectors by the length of the k_1 keyword subvector in the case where keywords were used as the terms in causal

[36] When binary weighting is used, the degree of similarity between a query and a document is based on the percentage of query terms that appear in the document. When term frequency weighting is used, the degree of similarity between a query and a document is based on the number of times query terms appear in the document (with different weights for different query terms).

[37] Actually, this is true only for *ad hoc* queries experiments where the same set of weights for combining subscores is used irrespective of the query. Normalization of query vectors has no effect in *routing queries* experiments where the set of weights to use is determined separately for each query.

relations, and the r_1 Roget code subvector in the case where Roget codes were used.[38] In effect, the length of the term subvector was taken as an approximation of the length of the *relation* subvector (which would include not just the causal relation but all the other relations in the query).[39]

5.3.2.5. Similarity measure

This section describes how the degree of match between the query and a document was calculated. The calculation involved two steps:

1. Calculate the similarity between each query subvector and the corresponding document subvector as listed in 5. The similarity measure used was the *inner product* measure:
   $\text{SIMILARITY(query, document)} = \Sigma^{n}_{k=1} [\ \text{WEIGHT}_{\text{query}}(\text{k}) * \text{WEIGHT}_{\text{doc}}(\text{k})\ ]$
   where $\text{WEIGHT}_{\text{query}}(\text{k})$ denotes the weight for component k in the query vector.
   $\text{WEIGHT}_{\text{doc}}(\text{k})$ denotes the weight for component k in the document vector.
   n is the number of components in the vectors.
   A similarity subscore was thus generated for each query subvector.

2. Combine the similarity subscores generated in Step 1 into a composite similarity score by taking the weighted sum of the similarity subscores.

In step 2, the appropriate set of weights to use when taking the weighted sum of similarity subscores was determined empirically. For each of the query subvectors listed in 5, a weight had to be assigned to it. For the ad hoc queries experiment, I had to determine one set of weights to be used for all the queries. For the routing queries experiment, the weights were customized for each query. For each query, I had to determine the best set of weights to use.

For the ad hoc queries experiment, the best set of weights was determined empirically by trying out various weights and finding the set of weights that gave the best retrieval results using the *normalized recall* and *normalized precision* measures (described in the next section). The formula used to combine the subscores from the different subvectors was of the form:

$$\text{k}_1/(\max \text{k}_1) + \text{w}_1 * \text{k}_2/(\max \text{k}_2) + \text{w}_2 * \text{r}_1/(\max \text{r}_1) + \text{w}_3 * \text{r}_2/(\max \text{r}_2) + ...$$

where k_1, k_2, r_1, r_2 ... were the subscores and w_i were the weights to be determined. *max* k_1, *max* k_2 ... refer to the highest subscore obtained for the particular subvector by a document. The purpose of dividing each subscore by the highest value was simply to rescale the subscores to be in the range [0,1]. For each subscore, the square and square root of the subscore were also tried. For example, the squared term k_1^{2} and the square root term $\text{k}_1^{1/2}$ were also tried in place of the linear term k_1. Interaction terms k_1*r_1, k_1*r_2, k_1*v, k_1*vp ... (i.e. interaction between the keyword subscore with the other subscores) were also explored.

For the routing queries experiment, it was obviously not possible to determine manually the best set of weights for each query. Stepwise logistic regression (using the function *PROC*

[38] It is more appropriate to use the length of the k_2 and r_2 subvectors for "normalizing" the causal relation subvectors, because k_2 and r_2 use binary weighting as is the case with all the causal relation subvectors used in this study. The reason this wasn't done was because the k_2 and r_2 subvectors were added to the study after all the other subvectors had been constructed. I felt that it wasn't worth the effort to recalculate all the causal relation subvectors using the length of the k_2 and r_2 subvectors.

[39] Preliminary experiments indicated that the use of unnormalized relation subvectors does not perform as well as the normalization method used here, supporting the notion that this approximation is helpful.

LOGISTIC in SAS ver. 6) was used instead to determine the best set of weights to use for each query. In the logistic regression, the documents were treated as the experimental units. The independent variables were the subscores for the different subvectors, and the dependent variable was the relevance of the document (whether the document was relevant or not relevant to the query). In the stepwise logistic regression, $\alpha=0.10$ was the threshold used for selecting variables to enter the regression model and $\alpha=0.15$ was the threshold used for eliminating variables from the model. The logistic regression would find a set of weights for the independent variables that would best predict the probability that a document was relevant (Hosmer & Lemeshow, 1989). The weights determined by logistic regression using one set of documents was then applied to the new set of documents, and the documents ranked in decreasing value of the estimated probability of relevance.[40]

The advantage of using logistic regression is that a large number of variables can be tried using stepwise regression. The disadvantage is that logistic regression does not directly optimize the retrieval measures (*normalized recall* measure, *normalized precision*, *11-point recall-precision average*, and *3-point recall-precision average*) which are based on the ranks of the documents. It is possible to obtain a significantly better model as determined by logistic regression but which does not yield an improvement in the retrieval effectiveness measures based on the ranks of documents.

5.3.3. Retrieval effectiveness measures

The retrieval effectiveness measures (the dependent variable) used in this study were the *normalized recall*, the *normalized precision*, the *11-point recall-precision average* and the *3-point recall-precision average* (Salton & McGill, 1983, pp. 180-182). These measures assume that the retrieval system produces a ranked list of documents with the document most likely to be relevant appearing first. They also assume that the relevance judgments are dichotomous, i.e. each document is judged to be either relevant or not relevant.

The *normalized recall* is given by the formula:

$$\text{normalized recall} = 1 - \frac{\Sigma^{REL}_{i=1}RANK_i - \Sigma^{REL}_{i=1}i}{REL * (N - REL)}$$

where *REL* is the number of relevant documents, $RANK_i$ represents the rank of the *i*th relevant document, and *N* is the total number of documents in the database. The measure reflects the number of ranks the relevant documents must move up in order to obtain perfect performance (with all the relevant documents ranked above the non-relevant documents). The *normalized recall* measure is equivalent to *the percentage of concordant pairs*, a measure of association that can be produced by many statistical packages.[41] Salton and McGill (1983) noted that this measure is sensitive to the rank

[40] Actually, the documents were ranked in decreasing value of their estimated *logits*, a *logit* being the *log* of the odds that the document was relevant.

[41] The percentage of concordant pairs is calculated as follows. Pair each *relevant* document with each *non-relevant* document. If *N* is the number of documents in the document collection and *REL* is the number of documents that are relevant to the query, then *REL * (N-REL)* is the total number of pairs that are formed. A pair is said to be *concordant* if the *relevant* document in the pair has a higher score than the *non-relevant* document. A pair is *discordant* if the *non-relevant* document has a higher score than the *relevant* document.

Assuming that there are no tied pairs, the *normalized recall* is the same as the *percentage of concordant pairs*:

$$\text{normalized recall} = 1 - \frac{\Sigma^{REL}_{i=1}RANK_i - \Sigma^{REL}_{i=1}i}{REL * (N - REL)}$$

of the last relevant document in the ranked list of documents output by a system.

The *normalized precision* is calculated using the formula:

$$\text{normalized precision} = 1 - \frac{\Sigma^{REL}_{i=1}\log \text{RANK}_i - \Sigma^{REL}_{i=1}\log i}{\log [\ N! \,/\, ((N\text{-}REL)!\ REL!)\ ]}$$

The *normalized precision* measure is similar to the *normalized recall* measure but places more emphasis on the higher ranked documents.

The *11-point recall-precision average* and the *3-point recall-precision average* are calculated using the interpolated precision versus recall graph, described in Salton and McGill (1983, p. 167-168). The precision[42] of the retrieved set of documents at 0.0, 0.1, 0.2, . . . , 0.8, 0.9 and 1.0 recall are determined from the graph. The *11-point recall-precision average* is the average precision at the 11 recall points. The *3-point recall-precision average* is the average of the precision at 3 recall points -- 0.2, 0.5 and 0.8. The *11-point average* and the *3-point average* summarize the recall-precision curve for the query. The *3-point average*, in contrast to the *11-point average*, ignores the extreme ends of the precision-versus-recall curve.

5.4. The Experiments

5.4.1. Ad hoc queries experiment

The *ad hoc queries experiment* was carried out using 39 query statements for model building (i.e. for determining the best set of weights to use when combining the subscores from the different types of matching) and using a different set of 38 query statements for model validation (i.e. for determining the retrieval effectiveness of the models developed). I shall refer to the first set of 39 queries as the *training set*, and the second set as the *testing set*.

$$= \quad 1 - \frac{\text{no. of discordant pairs}}{\text{total no. of pairs}}$$

$$= \quad 1 - \text{proportion of discordant pairs}$$

$$= \quad \text{proportion of concordant pairs}$$

[42] *Precision* is defined as the percentage of the retrieved set of documents that are judged to be relevant by the user. *Recall* is the percentage of all the documents that are relevant to the user which are actually in the retrieved set of documents.

Model No.	Model
model 0	0.0 *documents are ranked in the order they appear in the database*
model 1	k_1 *keyword matching using term frequency weighting*
model 2	$k_1/(\max k_1) + 0.62*k_2/(\max k_2)$ *keyword matching using a combination of term frequency weighting and binary weighting*
model 3	$k_1/(\max k_1) + 0.62*k_2/(\max k_2) + 0.70*v^2/(\max v^2) + 0.65*c/(\max c)$ *combination of keyword matching and causal relation matching*
model 4	$k_1/(\max k_1) + 0.62*k_2/(\max k_2) + 0.87*a^{1/2}/(\max a^{1/2})$ *combination of keyword matching and word proximity matching*
model 5	$k_1/(\max k_1) + 0.62*k_2/(\max k_2) + 0.87*a^{1/2}/(\max a^{1/2}) + 1.24*v^2/(\max v^2)$ *combination of keyword matching, word proximity matching and causal relation matching*
model 1b	$k_1/(\max k_1) - 0.35*r_1/(\max r_1)$ *combination of keyword matching and Roget code matching*

Table 6. Models developed in the ad hoc queries experiment

The models developed using the training set of queries are given in 5. Model 1, 2, 3, 4 and 5 use stemmed keywords as terms, whereas Model 1b uses a combination of stemmed keywords and Roget category codes. Model 1 represents keyword matching using term frequency weighting. Model 2 represents keyword matching using a combination of term frequency weighting and binary weighting. The retrieval results for Model 2 are taken as the baseline keyword matching results, against which the effect of using causal relation matching is assessed. Model 3 uses a combination of keyword matching and causal relation matching. Of the several types of causal relation matching tried, only "word->*" matching and "word->word" matching produced an improvement in the average retrieval results for the training set of queries. Model 4 uses keyword matching plus word proximity matching (i.e. matching pairs of words that co-occur within document sentences with pairs of causally related words in the query statement). Model 5 uses keyword matching plus word proximity matching plus causal relation matching.

Model 1b represents keyword matching using both stemmed keywords and Roget category codes. Term frequency weighting is used for both keyword matching and for Roget code matching. I attempted to build Model 2b using k_1, k_2, r_1 and r_2 subscores (i.e. using term frequency weighting as well as binary weighting for both keywords and Roget codes), but found that the r_2 subscore didn't improve retrieval when it was added to k_1, k_2 and r_1 subscores.

model 0 documents are ranked in the order they appear in the database (equivalent to retrieving documents randomly)
model 1 keyword matching using term frequency weighting
model 2 keyword matching using a combination of term frequency weighting and binary weighting (baseline)
model 3 combination of keyword matching and causal relation matching
model 4 combination of keyword matching and word proximity matching
model 5 combination of keyword matching, word proximity matching and causal relation matching
model 1b combination of keyword matching and Roget code matching

	model 0	**model 1**	**model 2**	**model 3**	**model 4**	**model 5**	**model 1b**
Precision at 11 recall levels							
0%	0.4500	0.8431	0.8562	0.8877	0.8915	0.9049	0.8833
10%	0.3108	0.7116	0.7226	0.7436	0.7707	0.7657	0.7305
20%	0.2972	0.6472	0.6568	0.6755	0.6765	0.6919	0.6564
30%	0.2888	0.5881	0.6050	0.6301	0.6240	0.6350	0.6078
40%	0.2839	0.5547	0.5679	0.5818	0.5767	0.5776	0.5583
50%	0.2810	0.5197	0.5429	0.5403	0.5511	0.5477	0.5181
60%	0.2780	0.4736	0.4986	0.5012	0.5027	0.4958	0.4696
70%	0.2727	0.4313	0.4449	0.4425	0.4505	0.4497	0.4292
80%	0.2685	0.4035	0.4159	0.4153	0.4211	0.4250	0.4042
90%	0.2635	0.3704	0.3765	0.3801	0.3835	0.3859	0.3711
100%	0.2373	0.3043	0.3041	0.3046	0.3046	0.3045	0.3019
Average precision over the 11 recall levels							
	0.2938	0.5316	0.5447	0.5548	0.5594	0.5621	0.5391
Average precision over 3 recall levels (20%, 50% and 80% recall)							
	0.2822	0.5235	0.5385	0.5432	0.5496	0.5549	0.5262
Normalized recall							
	0.5411	0.7790	0.7909	0.7933	0.8008	0.8026	0.7811
Normalized precision							
	0.4130	0.6744	0.6893	0.6967	0.7032	0.7045	0.6817

Table 7. Retrieval results obtained when the models are applied to the training set of queries used to develop the models

	model 0	**model 1**	**model 2**	**model 3**	**model 4**	**model 5**	**model 1b**
Precision at 11 recall levels							
0%	0.2765	0.7670	0.7934	0.8035	0.7615	0.7505	0.7565
10%	0.2469	0.6150	0.6491	0.6414	0.6616	0.6607	0.6277
20%	0.2320	0.5663	0.6095	0.6059	0.6042	0.5878	0.5827
30%	0.2274	0.4927	0.5415	0.5451	0.5332	0.5435	0.5023
40%	0.2234	0.4529	0.4840	0.4865	0.4812	0.4867	0.4611
50%	0.2179	0.4238	0.4480	0.4446	0.4419	0.4450	0.4225
60%	0.2161	0.3897	0.4118	0.4183	0.4122	0.4116	0.3871
70%	0.2120	0.3624	0.3848	0.3869	0.3813	0.3810	0.3557
80%	0.2088	0.3300	0.3510	0.3476	0.3527	0.3525	0.3225
90%	0.2054	0.2957	0.3244	0.3231	0.3172	0.3152	0.2913
100%	0.1989	0.2450	0.2678	0.2669	0.2633	0.2625	0.2442
Average precision over the 11 recall levels							
	0.2241	0.4491	0.4787	0.4791	0.4737	0.4725	0.4503
Average precision over 3 recall levels (20%, 50% and 80% recall)							
	0.2196	0.4400	0.4695	0.4660	0.4663	0.4618	0.4426
Normalized recall							
	0.5422	0.7540	0.7705	0.7674	0.7626	0.7613	0.7553
Normalized precision							
	0.3594	0.6156	0.6382	0.6361	0.6309	0.6300	0.6184

Table 8. Retrieval results obtained when the models were applied to a new set of queries

Using Roget codes in causal relations also did not improve retrieval results. I attempted to add "Roget code->*", "*->Roget code" and "Roget code->Roget code" matching to Model 1b, but could obtain no improvement in retrieval. Using causal relations between Roget codes *in addition to* causal relations between keywords also did not improve retrieval.

The retrieval results obtained by applying the models to the training set of queries are given in 7. Model 5 gave the best results, with an improvement of 3.0% in the *3-point recall-precision average* over Model 2.[43] Using Roget code matching in addition to keyword matching (Model 1b) yielded an improvement of only 0.5% in the *3-point average* (compared with Model 1). The improvement is slightly greater for the *11-point average* -- 1.4% improvement in the *11-point average* between Model 1b and Model 1.

[43] Of the four summary retrieval measures, the *11-point average*, the *3-point average*, *normalized recall* and *normalized precision*, I found the *3-point average* to be the most sensitive measure. I shall use mainly the *3-point average* in reporting retrieval results.

Figure 4. Recall-precision curve for Model 2 and Model 5 in the *ad hoc queries* experiment

7 gives the retrieval results obtained when the models are applied to the testing set of 38 queries. Comparing the results for Model 2 and Model 5 in 7 and their recall-precision curves shown in 4, it is clear that causal relation matching did not produce any improvement in retrieval. In fact, there is a small degradation in the *3-point average*. Comparing the results for Model 1 and Model 1b in 7 and their recall-precision curves shown in 4, we find a small and non-significant improvement in the precision at recall levels 10% to 40%. The only substantial improvement found in this experiment was between Model 1 and Model 2 (about 6.7% improvement in the *3-point average*). Using binary weighting in addition to term frequency weighting of keywords produced a clear improvement in retrieval effectiveness.

Figure 5. Recall-precision curves for Model 1 (keyword) and Model 1b (keyword plus Roget code) in the *ad hoc queries* experiment

In conclusion, causal relation matching was not found to improve retrieval effectiveness in the *ad hoc queries* experiment. In this experiment, the best set of weights to use for combining the subscores from the different types of matching was determined using one set of queries and tested using a different set of queries. The purpose was to develop one retrieval strategy that, though not optimal for any single query, will on average give good retrieval results across all the queries. The fact that no improvement in retrieval was found in this experiment does not necessarily mean that causal relation matching cannot be used to improve retrieval effectiveness. It may be that the different types of causal relation matching don't work equally well for all the queries, and each type of causal relation matching should be weighted differently for different queries. A particular type of causal relation matching may capture the similarity between document and query better for one query than another. The poor results from this experiment only show that it is not feasible to develop one strategy for using causal relation matching that can be applied uniformly to all the queries.

5.4.2. Routing queries experiment

The *routing queries* (or SDI) experiment was designed to test whether causal relation matching would improve retrieval results when the set of weights used for combining the subscores from the different types of matching was determined separately for each query. The set of weights to use for each query was determined using one half of the test database (using documents from Wall Street Journal 1986, 1987, and 1991), and then tested using the other half of the test database (using documents from Wall Street Journal 1988, 1989, 1990, and 1992). The weights were determined using forward stepwise logistic regression, using a threshold of 0.10 for selecting variables to enter the model and a threshold of 0.15 for eliminating variables from the model. Before a regression model was accepted, it was first checked by applying it to the first half of the test database used for developing the model. The model was accepted only if it didn't produce a lower *normalized recall* than the default model (which was Model 2 in the case of causal relation matching). For some queries, there were too few relevant documents to complete the stepwise regression. The default model was also used in these cases.

Eleven models were developed for each query: five models that made use of keyword stemming without Roget codes, and another six that used both stemmed keywords and Roget codes. I shall first describe the results for the five models not using Roget codes, and then describe the results when Roget codes were used in addition to keywords.

5.4.2.1. Routing queries experiment using keywords (without Roget codes)

Model No.	Variables Used in the Stepwise Regression
model 1	k_1
keyword matching using term frequency weighting	
model 2	k_1 k_2 k_1^2 k_2^2 $k_1^{1/2}$ $k_2^{1/2}$
keyword matching using a combination of term frequency weighting and binary weighting	
model 3	k_1 k_2 v vp w wp c cp k_1^2 k_2^2 v^2 vp^2 w^2 wp^2 c^2 cp^2 $k_1^{1/2}$ $k_2^{1/2}$ $v^{1/2}$ $vp^{1/2}$ $w^{1/2}$ $wp^{1/2}$ $c^{1/2}$ $cp^{1/2}$ k_1*v k_1*vp k_1*w k_1*wp k_1*c k_1*cp
combination of keyword matching and causal relation matching	
model 4	k_1 k_2 a ap k_1^2 k_2^2 a^2 ap^2 $k_1^{1/2}$ $k_2^{1/2}$ $a^{1/2}$ $ap^{1/2}$ k_1*a k_1*ap
combination of keyword matching and word proximity matching	
model 5	k_1 k_2 a ap v vp w wp c cp k_1^2 k_2^2 a^2 ap^2 v^2 vp^2 w^2 wp^2 c^2 cp^2 $k_1^{1/2}$ $k_2^{1/2}$ $a^{1/2}$ $ap^{1/2}$ $v^{1/2}$ $vp^{1/2}$ $w^{1/2}$ $wp^{1/2}$ $c^{1/2}$ $cp^{1/2}$ k_1*a k_1*ap k_1*v k_1*vp k_1*w k_1*wp k_1*c k_1*cp
combination of keyword matching, word proximity matching and causal relation matching	

Table 9. Variables entered into the stepwise logistic regression for 5 models not using Roget codes

The variables (i.e. the subscores from the different types of matching) that were used in the logistic regression to develop the five models are listed in 9. The retrieval results for the five models are given in 9. Model 2 represents keyword matching using a combination of term frequency weighting and binary weighting. The results for Model 2 are taken as the baseline results. The results for Model 1 (keyword matching using term frequency weighting alone) are given to show how Model 2 compares with it. Model 3 uses causal relation matching, Model 4 uses word proximity matching, and Model 5 uses both word proximity matching and causal relation matching. The comparisons of interest are between Model 3, 4 and 5 with the baseline of Model 2. Dunnett's *t* test (one-tailed) was used to test whether each of the Model 3 to 5 gave better results than Model 2.

model 1 keyword matching using term frequency weighting (k_1 subvector)
model 2 keyword matching using a combination of term frequency weighting and binary weighting (combining the scores from k_1 and k_2 subvectors)
model 3 combination of keyword matching and causal relation matching
model 4 combination of keyword matching and word proximity matching
model 5 combination of keyword matching, word proximity matching and causal relation matching

Precision at 11 recall levels

	model 1	**model 2**	**model 3**	**model 4**	**model 5**
0%	0.7691	0.7929	0.7972	0.8063	0.8246
10%	0.6656	0.6982	0.7116	0.7143	0.7284
20%	0.5884	0.6150	0.6458**	0.6375*	0.6521**
30%	0.5277	0.5457	0.5736*	0.5698	0.5918***
40%	0.4928	0.5238	0.5404	0.5386	0.5517**
50%	0.4681	0.4957	0.5084	0.5134*	0.5201**
60%	0.4344	0.4549	0.4613	0.4612	0.4708
70%	0.4138	0.4162	0.4177	0.4225	0.4268
80%	0.3871	0.3877	0.3875	0.3909	0.3944
90%	0.3522	0.3449	0.3423	0.3464	0.3477
100%	0.3073	0.2975	0.2923	0.2959	0.2891

Average precision over the 11 recall levels

	model 1	**model 2**	**model 3**	**model 4**	**model 5**
	0.4915	0.5066	0.5162	0.5179	0.5270**

Average precision over 3 recall levels (20%, 50% and 80% recall)

	model 1	**model 2**	**model 3**	**model 4**	**model 5**
	0.4812	0.4995	0.5139*	0.5139*	0.5222**

Normalized recall

	model 1	**model 2**	**model 3**	**model 4**	**model 5**
	0.7713	0.7875	0.7838	0.7888	0.7864

Normalized precision

	model 1	**model 2**	**model 3**	**model 4**	**model 5**
	0.6378	0.6652	0.6692	0.6724	0.6775

Dunnett's *t* test (1-tailed) was performed to compare the retrieval results for models 3, 4 and 5 with model 2, which was taken as the baseline. An asterisk indicates that the result is significantly better:

* $p<0.05$
** $p<0.01$
*** $p<0.001$

Table 10. Retrieval results from the routing queries experiment without the use of Roget codes

If we examine the precision figures at the various recall levels (9 and 6), we find that Model 3 gave better precision than Model 2 at all recall levels less than 80%, and was significantly better at 20% and 30% recall. Model 4 and Model 5 were better than Model 2 at all recall levels less than 100% (see 6). Model 5 was significantly better (*p*<0.01) at recall levels 20% to 50%.

Using the *3-point recall-precision average* as the retrieval effectiveness measure, we find that Model 3, 4 and 5 all did significantly better than Model 2 at the α=0.05 level. Model 3 and Model 4 each obtained a retrieval improvement of 2.9% in the *3-point average* compared with Model 2. Model 5 (which uses both word proximity matching and causal relation matching) obtained a retrieval improvement of 4.5% over Model 2, and this improvement was significant at the α=0.01 level. The *normalized recall* measure did not show any improvement for Model 3, 4 and 5 compared with Model 2. It is known that the *normalized recall* measure is sensitive to the ranks of the lowest ranked documents (Salton & McGill, 1983). For *normalized precision*, which places more emphasis on the ranks of the higher-ranked documents, we see that Model 3, 4 and 5 fared better than Model 2, but the improvements were not significant.

Subsidiary question 2.4a asks whether causal relation matching gives better retrieval results than term proximity matching. Comparing the results for Model 3 and 4 in 9, we find that Model 4 (using term proximity matching) fared slightly better than Model 3 (using causal

Figure 6. Recall-precision curves for Model 2 and Model 3 in the *routing queries* experiment.

Results averaged over 41 queries

	model 1	**model 2**	**model 3**	**model 4**	**model 5**
Average precision over the 11 recall levels	0.5901	0.6061	0.6206	0.6277*	0.6375**
Average precision over 3 recall levels (20%, 50% and 80% recall)	0.5789	0.5972	0.6192*	0.6225**	0.6329***
Normalized recall	0.7761	0.7914	0.7912	0.8021*	0.8028*
Normalized precision	0.6998	0.7212	0.7303	0.7388*	0.7480**

Table 11. Results from SDI experiments for queries with more than 50 relevant documents in database

Figure 7. Recall-precision curves for Model 2 and Model 5 in the *routing queries* experiment

relation matching). This suggests that, for the purpose of information retrieval, if a query asks for documents that indicate that *A* causes/caused *B*, we can assume that a causal relation exists between *term A* and *term B* in a document simply from the fact that the two terms occur within the same sentence.

Although causal relation matching did not give better retrieval results than term

Centrality of the causal relation	Model 1	Model 2	Model 3	Model 4	Model 5
Non-central (N=24)	0.4533	0.4664	0.4736	0.4681	0.4748
Central (N=48)	0.4952	0.5160	0.5341	0.5368	0.5459

Table 12. *3-point recall-precision averages* for two groups of queries: one group for which the causal relation is peripheral and the second group of queries for which the causal relation is central

proximity matching, if we use both causal relation matching *as well as* term proximity matching, the causal relation matching might yield some improvement in retrieval effectiveness over and above any improvement from the term proximity matching. *Subsidiary question 2.4b* asks whether using causal relation matching *in addition to* term proximity matching will give better results than using term proximity matching alone. The recall-precision figures (9) indicate that Model 5 did better than Model 4 at all recall levels less than the 100% level. However, the improvements were not significant.

It is not clear why the precision at 100% recall got progressively worse from Model 1 to Model 5. It may be an artifact of the way the test databases were constructed. The test database for each query was constructed from the union of the top-ranked 100 documents retrieved by each system that participated in TREC-1 and TREC-2 conferences. (Most of the systems used some variation of keyword matching.) So, there was a cut-off point for including documents in the test database. Since the documents below the cut-off point were not included in the test database, these documents were effectively prevented from moving up in ranks above the bottom-ranked documents in the test database. This might have caused the precision near the 100% recall level to be artificially depressed.

The size of a test database has an effect on how good a regression model is in estimating the likelihood that a document is relevant. How good a regression model is depends on the sample size that is used to develop it. In typical retrieval situations, the number of relevant documents is much smaller than the number of non-relevant documents. So, how good a regression model is depends on the number of relevant documents in the sample. In this experiment, I would expect causal relation matching to do better for queries which have a higher number of relevant documents in the test database. I found this to be the case. 11 summarizes the retrieval results for the 41 queries that have more than 50 relevant documents in the test database. The average improvement in the *3-point average* for these 41 queries for Model 5 was 6.0% (compared to 4.5% improvement for all the 72 queries). Also, Model 4 and 5 did significantly better than Model 2 for all the four retrieval measures (*11-point average*, *3-point average*, *normalized recall* and *normalized precision*).

Subsidiary question 2.3a asks whether there is a greater improvement in retrieval results from causal relation matching when the causal relation is *central* to the user's query than when the causal relation is *peripheral* to the query. To answer this question, the retrieval results were reanalyzed as a split-plots factorial design with *centrality* (whether the causal relation is central to the query or not) as the between subjects factor. In the analysis of variance, *centrality* was not found to be significant. Neither was the interaction term *centrality*type of matching*. So the improvement in retrieval results was not significantly different when the causal relation was central to the query than when it was not. The *3-point averages* for the two groups of queries (central and non-central) are given in 12. For the queries in which the causal relation is central, Model 5 obtained an improvement of 5.8% over Model 2. In contrast, Model 5 obtained an improvement of only 1.8% for the queries in which the causal relation is peripheral. As I mentioned earlier, the difference was not found to be significant.

The *subsidiary questions 2.3b, 2.3c and 2.3d* ask whether the improvement in retrieval

Improvement in 3-point average	Causal Association	Result from word matching	Avg. causal relation score	
			per document	per relevant doc
1. Amount of improvement	0.0542	-0.1480	-0.0976	-0.0964
2. % improvement over baseline	0.1263	-0.0849	-0.1663	-0.3141
3. % of possible improvement	-0.0840	-0.2283*	-0.1473	-0.0600

Table 13. Correlations between the improvement in *3-point average* for Model 3 and four variables

effectiveness from causal relation matching is greater:

- when the *causal association* between the two terms in the relation is weak than when the association is strong (subsidiary question 2.3b)
- when the query has poor retrieval results from term matching than when the query has good results from term matching (subsidiary question 2.3c)
- when many causal relation matches are found during retrieval than when there are few causal relation matches (subsidiary question 2.3d).

In the analysis carried out to answer the 3 subsidiary questions, the improvement in retrieval effectiveness from causal relation matching is taken to be the difference in the *3-point average* between Model 3 and Model 2. The *improvement in retrieval effectiveness* can be expressed in terms of

1. the amount of improvement, i.e. the difference in the *3-point average*
2. percentage improvement over the baseline result, i.e.

   $$\frac{\text{the difference in the } \textit{3-point average}}{\text{the } \textit{3-point average} \text{ for the baseline model}} \quad * \quad 100$$

   This is the measure that is commonly used in the *information retrieval* literature. I use mainly this measure elsewhere in this report.
3. percentage of the possible improvement in the baseline result, i.e.

   $$\frac{\text{the difference in the } \textit{3-point average}}{1.0 - \text{the } \textit{3-point average} \text{ for the baseline model}} \quad * \quad 100$$

I shall refer to these as *improvement measures* (1), (2) and (3).[44] To answer the three subsidiary

[44] Measure (1), *amount of improvement*, places more emphasis on the retrieval improvement of the queries that have poor results with the baseline model. If a query already has good results with the baseline model, there is not so much room for improvement.

Measure (2), *percentage over the baseline result*, places even more emphasis on the queries that do poorly with the baseline model. Since the amount of improvement is divided by the baseline result, the smaller the baseline result, the higher the percentage is.

Measure (3), *percentage of possible improvement*, does not place so much emphasis on the queries that have poor results with the baseline model. Whereas measure (2) focuses on the *relevant* documents, measure (3) focuses on the *non-relevant* documents. Measure (2) measures the percent increase in the proportion of *relevant* documents at each recall level. Measure (3) measures the percent decrease in the proportion of *non-relevant* documents at each recall level.

% Improvement	No. of Queries	Query Nos.
50% to 100%	2	75,146
30% to 50%	4	7,120,122,130
20% to 30%	7	14,17,19,51,60,68,145
10% to 20%	5	5,26,90,106,137
5% to 10%	0	
>0% to 5%	4	1,22,58,124
0%	30	
<0% to -5%	9	
-5% to -10%	4	
-10% to -20%	2	71,72
-20% to -30%	3	15,69,105
-30% to -50%	2	25,141

Table 14. Percent improvement in 3-point recall-precision average for model 5 compared with model 2

research questions, I calculated the product moment correlation (Pearson *r*) between the three improvement measures and the following four variables:

1. the strength of the causal association
2. the *3-point average* for Model 2 (i.e. the retrieval result for keyword matching)
3. average *causal relation matching* score per document
4. average *causal relation matching* score per *relevant* document.

The four variables have been explained in Section 5.2.4. The correlations are given in 13.

For causal association, I expected the correlation to have a negative value, i.e. the weaker the causal association, the greater the retrieval improvement from causal relation matching. We can see in 13, that only the correlation with *improvement measure* (3) has the expected negative sign. This correlation is not significantly less than 0.

For the correlation with the keyword matching result, I also expected the correlation to have a negative value, i.e. the poorer the result from keyword matching, the greater the improvement from causal relation matching. 13 shows this to be the case for all three improvement measures. Only the correlation with *improvement measure* (3) (percentage of possible improvement) is significantly less than 0 (p<0.05 for a 1-tailed test).

I turn now to the correlations between the 3 improvement measures and the average causal relation score *per document* as well as the average causal relation score *per relevant document* (columns 3 and 4 in 13). I expected the correlations to have positive values, but 13 shows that the correlations are negative. The negative correlations suggest that if there are many causal relation matches in the database, the query is less likely to benefit from causal relation matching. I can think of two reasons why this may be so:

1. The high number of causal relation matches may be due to a particular causal relation occurring frequently in the database. A causal relation that occurs frequently in the database is less likely to be helpful in distinguishing relevant documents from non-relevant ones. In this study, the causal relation matches were weighted by the inverse document frequency (*idf*) which lowered the score for common causal relations. However, the *idf* values were calculated for the whole of the Wall Street Journal collection and not for the test database for each query.

2. The high number of causal relation matches may be due to two query terms (call them *term A* and *term B*) that almost always have a causal relation between them whenever the two terms co-occur in a document. In such a case, the mere presence of the two terms in a document already signal a causal relation between them, and "*term A->term B*" relation matching cannot do better than plain keyword matching in identifying relevant documents.

Let us now look more closely at the individual queries. Of the 72 queries, 22 queries (31%) had better retrieval results (measured by the *3-point average*) with Model 5 (word proximity plus causal relation matching) than with Model 2 (baseline keyword matching model). The percentage improvement for individual queries is summarized in 14. As can be seen in the table, 18 of the queries had improvement of over 10%. On the other hand, 7 queries suffered a drop of more than 10% in the retrieval result. For these 7 queries, Model 5 improved the retrieval results for the first half of the test database used in developing the model, but not for the second half of the database. It appears that the regression models developed for these 7 queries do not have much predictive power.

% Improvement	No. of Queries	Query Nos.
50% to 100%	2	75,120
30% to 50%	1	7
20% to 30%	1	14
10% to 20%	5	17,26,90,137,145
5% to 10%	1	146
>0% to 5%	7	1,5,22,58,110,124,130
0%	35	
<0% to -5%	12	
-5% to -10%	3	
-10% to -20%	2	69,122
-20% to -30%	1	15
-30% to -50%	2	25,141

Table 15. Percent improvement in 3-point recall-precision average for model 5 compared with model 4

Comparing Model 5 (word proximity plus causal relation matching) with Model 4 (word proximity matching alone), 17 queries (24%) had better results (measured by the *3-point average*) with Model 5 than with Model 4. For these queries, causal relation matching produced an improvement in retrieval results over and above any improvement from word proximity matching. The percentage improvement for individual queries is summarized in 15.[45] 9 queries had an

[45] For query 110, Model 5 produced a better retrieval result than Model 4, but not a better result than Model 2. This is because the result for Model 4 was worse than for Model 2.

improvement of over 10%.

% Improvement	No. of Queries	Query Nos.
50% to 100%	2	122,146
30% to 50%	1	130
20% to 30%	5	19,60,68,51,106
10% to 20%	3	5,25,145
5% to 10%	0	
>0% to 5%	4	1,17,22,135
0%	49	
<0% to -5%	3	
-5% to -10%	1	
-10% to -20%	2	71,72
-20% to -30%	1	105
-30% to -50%	1	120

Table 16. Percent improvement in 3-point recall-precision average for model 4 (using word proximity matching) compared with model 2

% Improvement	No. of Queries	Query Nos.
50% to 100%	2	122,146
30% to 50%	1	130
20% to 30%	5	19,60,68,51,106
10% to 20%	3	5,25,145
5% to 10%	0	
>0% to 5%	4	1,17,22,135
0%	49	
<0% to -5%	3	
-5% to -10%	1	
-10% to -20%	2	71,72
-20% to -30%	1	105
-30% to -50%	1	120

Table Error! Main Document Only.. Percent improvement in 3-point recall-precision average for model 4 (using word proximity matching) compared with model 2

Type of Causal Relation Matching	No. of Queries	Query Nos.
word -> *	6	1,5,14,75,130,145
word pair -> *	1	120
* -> word	9	5,7,14,17,26,58,124,130,146
* -> word pair	4	7,75,90,137
word -> word	3	22,124,137

Table 17. The number of queries for which each type of causal relation matching appears in the regression model for model 5

Comparing Model 4 (word proximity matching) with Model 2 (baseline model), 15 queries (21%) had better results with Model 4 than with Model 2. The percentage improvement for individual queries is summarized in 15.

Several types of causal relation matching were used in this study. The regression models for the queries that showed an improvement with Model 5 compared with Model 2 were examined to see which types of causal relation matching scores appeared in the regression models. 17 shows, for each type of causal relation matching, how many queries had the variable for that type of matching in the regression model for Model 5. From the table, it is clear that the most useful types of causal relation matching are those involving a wildcard, i.e. where one member of the relation -- either the *cause* or the *effect* -- is allowed to matching with anything. "Word->word" matching appears to have helped only three queries. I conclude that "word->word" matching generally does not improve retrieval effectiveness over and above word proximity matching. Word proximity matching should be used in place of "word->word" matching. On the other hand, causal relation matching with a wildcard as the *cause* or the *effect* can improve retrieval. 17 also shows that "*->word" and "*->word pair" matching helped nearly twice as many queries as "word->*" and "word pair->*" matching. This is probably because there are 15 queries which specify the *effect* but not the *cause*. These queries expect the *cause* to be supplied by the retrieved document. Two examples of such queries are given below:

> Query 7: A relevant document will cite at least one way to reduce the U.S. budget deficit.
>
> Query 40: A relevant document will give at least one reason why a particular U.S. Savings and Loan has failed, or is about to be closed by public authorities or acquired by another financial institution.

For such queries, the "word->*" and "word->word" matching cannot help. On the other hand, there is only 1 query which specifies the *cause* but not the *effect*.

Note: A blank indicates that the regression coefficient is 0.

% improvement		the sign of the regression coefficient for					
over Model 4	query no.	k_1*v	k_1*vp	k_1*w	k_1*wp	k_1*c	k_1*cp
98%	75		+				
96%	120						
30%	7					+	
22%	14		-				
20%	17						
18%	137					-	+
16%	90						
15%	26				+		
11%	145						
7%	146						

Table 18. The sign of the regression coefficients for six *keyword matching*causal relation matching* interaction terms for each of 10 queries

Subsidiary question 2.2a asks whether there is an interaction between the term matching score and the causal relation matching score in determining the relevance of a document. In effect, the question asks whether the interaction terms k_1*v, k_1*vp, k_1*w, k_1*wp, k_1*c, and k_1*cp (representing the interaction between keyword matching and the different types of causal relation matching) have non-zero coefficients in the regression model. The regression models for Model 5 were examined. Of the 10 queries that obtained more than a 5% retrieval improvement with causal relation matching over and above term proximity matching, 5 queries have non-zero coefficients for one or more of the interaction terms in their regression models. 18 lists these queries and the sign of the regression coefficient for each interaction term. A positive coefficient indicates that the higher the keyword matching score, the more effective is that particular type of causal relation matching. A negative coefficient indicates that the lower the keyword matching score, the more effective is that type of causal relation matching. Of the six interaction terms, no particular interaction term appears to predominate.

In conclusion, causal relation matching where either the *cause* or the *effect* is a wildcard can be used to improve retrieval effectiveness if the appropriate weight for each type of matching can be determined for each query -- as in an SDI or *routing queries* situation. The best results are obtained when causal relation matching is combined with word proximity matching. The retrieval improvement from causal relation matching is greater for queries that obtain poor results from keyword matching -- especially if the improvement is measured in terms of the percentage of the possible improvement. Causal relation matching also produces a bigger retrieval improvement when the causal relation is central to the query than when it is peripheral, but the difference is not significant.

5.4.2.2. Routing queries experiment with Roget codes

Model No.	Variables Used in the Stepwise Regression
Model 1b	k_1 r_1 $k_1{*}r_1$
a combination of keyword matching and Roget code matching using term frequency weighting	
Model 2b	variables in model 2 and r_1 r_2 r_1^2 r_2^2 $r_1^{1/2}$ $r_2^{1/2}$ $k_1{*}r_1$ $k_1{*}r_2$
a combination of keyword matching and Roget code matching using both term frequency weighting and binary weighting	
Model 3b	variables in model 3 and r_1 r_2 vr wr cr r_1^2 r_2^2 vr^2 wr^2 cr^2 $r_1^{1/2}$ $r_2^{1/2}$ $vr^{1/2}$ $wr^{1/2}$ $cr^{1/2}$ $k_1{*}r_1$ $k_1{*}r_2$ $k_1{*}vr$ $k_1{*}wr$ $k_1{*}cr$
a combination of term matching and causal relation matching using both keywords and Roget codes	
Model 4b	variables in model 4 and r_1 r_2 ar r_1^2 r_2^2 ar^2 $r_1^{1/2}$ $r_2^{1/2}$ $ar^{1/2}$ $k_1{*}r_1$ $k_1{*}r_2$ $k_1{*}ar$
a combination of term matching and term proximity matching using both keywords and Roget codes	
Model 5b	variables in model 5 and r_1 r_2 vr wr cr ar r_1^2 r_2^2 vr^2 wr^2 cr^2 ar^2 $r_1^{1/2}$ $r_2^{1/2}$ $vr^{1/2}$ $wr^{1/2}$ $cr^{1/2}$ $ar^{1/2}$ $k_1{*}r_1$ $k_1{*}r_2$ $k_1{*}vr$ $k_1{*}wr$ $k_1{*}cr$ $k_1{*}ar$
a combination of term matching, term proximity matching and causal relation matching using both keywords and Roget codes	
Model 5c	variables in model 5 and r_1 r_2 r_1^2 r_2^2 $r_1^{1/2}$ $r_2^{1/2}$ $k_1{*}r_1$ $k_1{*}r_2$
same as Model 5b, except that Roget codes are not used for term proximity matching and causal relation matching. Roget codes are used only for term matching.	

Table 19. Variables entered into the stepwise logistic regression for 5 models using Roget codes

In a preliminary experiment using 39 queries, I had already determined that using Roget code matching (instead of keyword matching) produced worse retrieval results than keyword matching for all the weighting schemes that were tried. The retrieval results for the preliminary experiment are given in Appendix 6.

The Models

model 1b keyword matching and Roget code matching, combining the scores from k_1 and r_1 subvectors

model 2b keyword matching and Roget code matching, combining the scores from k_1, k_2, r_1 and r_2 subvectors

model 3b combination of term matching and causal relation matching, using both keywords and Roget codes

model 4b combination of term matching and term proximity matching, using both keywords and Roget codes

model 5b combination of term matching, term proximity matching and causal relation matching, using both keywords and Roget codes

model 5c same as model 5b, except that Roget codes are not used for term proximity and causal relation matching. Roget codes are used only for term matching.

	model 1b	**model 2b**	**model 3b**	**model 4b**	**model 5b**	**model 5c**
Precision at 11 recall levels						
0%	0.7945	0.8422*	0.8392	0.8418	0.8330	0.8803
10%	0.6990	0.7378*	0.7225	0.7538	0.7365	0.7560
20%	0.6157	0.6525**	0.6522	0.6632	0.6588	0.6818
30%	0.5554*	0.5886***	0.5976	0.6101	0.6076	0.6252
40%	0.5251*	0.5580***	0.5560	0.5710	0.5644	0.5874
50%	0.4990*	0.5201***	0.5239	0.5399	0.5355	0.5508
60%	0.4628*	0.4746**	0.4806	0.4899	0.4922	0.5043
70%	0.4383*	0.4341*	0.4387	0.4418	0.4501	0.4558
80%	0.4127*	0.4055*	0.3981	0.4078	0.4053	0.4150
90%	0.3710*	0.3561*	0.3588	0.3613	0.3582	0.3685
100%	0.3169	0.3020	0.2996	0.3052	0.2962	0.3041
Average precision over the 11 recall levels						
	0.5173*	0.5338***	0.5334	0.5442	0.5398	0.5572
Average precision over 3 recall levels (20%, 50% and 80% recall)						
	0.5091*	0.5260***	0.5247	0.5370	0.5332	0.5492
Normalized recall						
	0.7875*	0.7989*	0.7916	0.7942	0.7927	0.7993
Normalized precision						
	0.6650**	0.6889***	0.6811	0.6887	0.6846	0.7011

1-tailed *t* test was carried out to compare the retrieval results for model 1b with those for model 1, and the results for model 2b with those for model 2. An asterisk indicates that the result is significantly better:

*	$p < 0.05$
**	$p < 0.01$
***	$p < 0.001$

Table 20. Retrieval results from the SDI experiment using Roget codes

% Improvement	No. of Queries	Query Nos.
50% to 100%	4	5,69,70,116
30% to 50%	8	14,25,75,90,103,106,131,145
20% to 30%	5	12,73,122,130,137
10% to 20%	5	7,19,23,24,68
5% to 10%	4	1,22,85,143
>0% to 5%	7	10,40,82,123,133,135,142
0%	25	
<0% to -5%	7	
-5% to -10%	1	149
-10% to -20%	4	43,45,125,141
-20% to -30%	1	104
-30% to -50%	1	51

Table 21. Percent improvement in 3-point recall-precision average for model 2b (keyword with Roget code matching) compared with model 2 (keyword matching only)

Figure 8. Recall-precision curves for Model 2 (keyword matching) and Model 2b (keyword plus Roget code matching)

In this section, I report the results of the *routing queries* experiment in which Roget codes were used *in addition to* keywords. The variables used in the stepwise logistic regression for the six models using both keywords and Roget codes are listed in 19. The retrieval results for the six models are summarized in 19.

Subsidiary question 2.1b asks whether the use of Roget codes *in addition to keywords* do improve retrieval effectiveness over using keywords alone. Comparing the results for Model 2b (keyword plus Roget code matching using both term frequency and binary weighting) in 19 with the results for Model 2 (keyword matching alone using term frequency and binary weighting) in 9, I found that Model 2b gave significantly better results than Model 2 at almost every recall level, as well as for the four summary measures (*11-point average*, *3-point average*, *normalized recall* and *normalized precision*). The recall-precision curves for Model 2 and Model 2b are given in 7. Model 1b (keyword plus Roget code matching using only term frequency weighting) also gave significantly better results than Model 1 (keyword matching using term frequency weighting). Using Roget code matching in addition to keyword matching clearly improved the retrieval results.

Model 2b produced an improvement of 5.3% in the *3-point average* over Model 2. Of the 72 queries, 33 queries (46%) showed an improvement in the *3-point average*. The percentage improvement obtained for individual queries is summarized in 19.

Subsidiary question 2.1c asks whether using Roget codes in addition to keywords in causal relation matching do improve retrieval effectiveness over using keywords alone in causal relation matching. The results show that using Roget codes as terms in causal relations did not improve retrieval results. This can be seen by comparing the results for Model 5b with the results for Model 5c in 19. Model 5b and Model 5c both use keywords and Roget codes for term matching. However, Model 5c uses only keywords for causal relation matching, whereas Model 5b uses both keywords and Roget codes for causal relation matching. Model 5b produced worse results than Model 5c.

Note: This table includes only queries with more than 5% improvement in *3-point average* for model 1b compared with model 1. A blank in the third or fourth column indicates that the regression coefficient is 0.

% improvement for		**the sign of the regression coefficient**	
over model 1	**query no.**	$\mathbf{r_1}$	$\mathbf{k_1*r_1}$
289%	131	+	-
215%	68	-	
200%	70	+	
83%	25	+	-
67%	5	-	-
53%	103	-	
53%	17	-	-
46%	75	-	
29%	15	-	
24%	19	-	-
22%	116		-
19%	146	-	-
16%	14	+	-
13%	1		-
12%	12		-
8%	23	+	-
7%	33	-	
6%	110	-	

Table 22. The sign of the regression coefficients for the Roget code matching variables in Model 1b

In the *ad hoc queries* experiment reported earlier, Model 1b, which was developed using 39 queries, had a negative weight for the Roget code matching subscore (see 5). This was unexpected. The negative weight indicated that the higher the score from Roget code matching, the less likely would the document be relevant. In other words, Roget code matching appeared to capture some *dissimilarity* between document and query not captured by keyword matching. The regression models for Model 1b in this *routing queries* experiment were examined to see whether the coefficients (or weights) for r_1 (the Roget code matching subscore) were positive or negative. 22 lists the queries that had more than a 5% improvement in the *3-point average* from Roget code matching. For each query, the sign of the coefficient for the r_1 variable as well as for the k_1*r_1 interaction term are given. The coefficient for r_1 is positive for 5 queries and negative for 10 queries. The *interaction term* appeared in the regression models for 11 queries. A negative coefficient for the *interaction term* indicates that Roget code matching is more helpful for documents with low keyword matching scores than for documents which already have high scores from keyword matching.

Is it possible to predict from the query statement whether the coefficient for r_1 is going to be positive or negative, and whether there is going to be an interaction between the keyword matching and Roget code matching scores? I shall leave this question to a future study. I shall only suggest a reason why the sign of the coefficient for r_1 is different for different queries.

I surmise that a Roget code match has a different effect on retrieval depending on whether the Roget code is for an important concept or a less important concept in the query statement. Not all the concepts in a query are equally important. A concept is important to a query if a document containing the concept has a high probability of being relevant. I surmise that the coefficient for r_1 will tend to be positive if:

1. an important concept in the query is expressed in the database using several synonymous or related terms, and
2. the synonyms or related terms share a common code in *Roget's International Thesaurus*.

If a concept in the query is always expressed using one particular word or phrase, then obviously replacing the word or phrase with a Roget code will not improve retrieval. If a concept has several synonyms but most of the synonyms are not listed in Roget's, then Roget code matching will also not be useful.

I surmise that the coefficient for r_1 will tend to be negative if there are several "unimportant" terms in the query statement. The unimportant terms tend to be common terms (i.e. they appear in a high percentage of documents) and common terms tend to have many synonyms. They also tend to have many senses, and hence many Roget codes. A document with many Roget code matches for the unimportant terms and hardly any Roget code matches for the important terms would be less likely to be relevant than a document with fewer matches for the unimportant terms and more matches for the important terms.

A quick look at 50 query statements reveal that for half of the queries, the important concepts either do not have synonyms or have synonyms that are not listed in *Roget's International Thesaurus* (3rd ed.). For the other half of the 50 queries, there are Roget codes only for some of the important concepts in each query. It appears that the main concepts of a query tend to be the less common concepts that are not likely to appear in a general purpose thesaurus like the Roget's. This may be why there are nearly twice as many queries with a negative coefficient for r_1 than a positive coefficient.

It is not easy to predict from a query statement whether the coefficient for r_1 is going to be positive or negative. I was not able to predict reliably, by reading the query statements, whether the coefficient for r_1 is positive or negative, and whether there is an interaction term in the regression model.

5.4.3. Qualitative analysis

This section reports some insights obtained by "eye-balling" sample documents for selected queries. I selected six queries that obtained some retrieval improvement with causal relation matching, and ten queries that didn't obtain any retrieval improvement with causal relation matching. For each of these queries, I selected for scrutiny two sets of documents:

1. five *non-relevant* documents that had the highest scores from causal relation matching.
2. five *relevant* documents that had no or few causal relation matches.

In the case of set 1 documents (*non-relevant* documents with one or more causal relation matches), I found that most of the causal relation matches were partial matches. For example, Query 7 asks for documents that state at least one way of reducing the U.S. budget deficit (i.e. *cause* the budget deficit to decrease). Some of the *non-relevant* documents had partial causal relation matches because the documents described what was causing the budget deficit *to increase*.

In the case of set 2 documents (*relevant* documents with no causal relation matches), there are several possible reasons why the causal relation of interest was not identified by the system:

1. the causal relation was not expressed explicitly in the document, but had to be inferred using world knowledge. In some cases, the *cause* concept or the *effect* concept was not even mentioned in the document, but could easily be inferred by the reader based on general knowledge.
2. the causal relation was explicitly expressed but there was no causal relation match because:
   a. the *cause* or the *effect* was expressed in the document using a different synonym than the terms used in the query statement
   b. the causal relation was expressed using a linguistic pattern or causal verb not known to the system (i.e. not in the set of linguistic patterns used in this study for identifying causal relations)
   c. the sentence structure was too complex for the system to process correctly.
3. the causal relation was not expressed in the document at all.

To obtain some idea of the relative importance of the above reasons, I counted the number of documents for which each of the above reasons held. I looked at sample relevant documents that did not have any *"word->*"* relation match, and then at sample relevant documents that did not have any *"*->word"* relation match.

In the case of *relevant* documents that did not have any *"word->*" matches*, Reason 1 (causal relation had to be inferred) was the most frequent reason. This reason accounted for 68% of the 25 documents (taken from 6 queries) examined.

In the case of *relevant* documents that did not have any *"*->word" matches*, Reason 1 (causal relation had to be inferred) accounted for 34% of the 38 documents (taken from 8 queries) examined. Reason 2b (the causal verb or the linguistic pattern used to indicate a causal relation is not known to the system) accounted for another 37% of the documents. Most of these linguistic patterns and causal verbs can have non-causal meanings in other contexts. Lexical disambiguation and contextual information is needed to interpret them correctly. Several of the instances involve nominalized causal verbs (causal verbs in noun form), for example:

(1) deficit *reduction*
(2) *expansion* of the theme park

Nominalized verbs are not handled in this study. It should also be pointed out that the computer program developed in this study was designed to identify causal relations in which both the cause and the effect are expressed. For retrieval purposes, it may be better to also identify instances of causal relations where the cause is not explicitly stated, as in the following sentences:

(3) John was killed. (i.e. *Someone* caused John to die.)

(4) Deficit reduction is the Administration's top priority. (i.e. The Administration hopes to reduce the deficit.)

(5) The expansion of the theme park is scheduled to begin next month. (i.e. *Some company* is expanding the theme park.)

5.5. Supplementary Analysis

This section reports the additional analysis that was carried out to address two questions. The first question relates to the queries which did not get better retrieval results with causal relation matching in the *routing queries* experiment. The second question relates to the queries which did get better retrieval results with causal relation matching.

In the case of the queries that *did not* get better retrieval results with causal relation matching, the lack of improvement may be due to errors in the automatic identification of causal relations. It may be that if the system could identify causal relations accurately, an improvement in retrieval results would have been found for these queries. So the question is:

> **Supplementary question 1**
> For the queries that didn't benefit from causal relation matching in the *routing queries* experiment, will causal relation matching yield an improvement in retrieval results if the automatic identification of causal relations in sentences is more accurate?

In the case of the queries that *did* get better results from causal relation matching, the improvement obtained from the causal relation matches might be obtainable simply by customizing the weights for individual keyword matches without resorting to causal relation matching. An example will make this clearer. Query 75 asks for documents that describe the effect of automation. In the *routing queries* experiment, the causal relation "automation->*" was found to be useful in improving the retrieval results. In the experiment, the appropriate weight to assign to a "automation->*" relation match was determined using logistic regression. (Since *binary weighting* was used for causal relation matches, the same weight was used even if "automation->*" occurred more than once in the document.) However, the appropriate weights to use for the individual keyword matches were not determined by the regression. Query 75 contains several other keywords besides "automation." During retrieval, a keyword matching score was calculated based on all the keyword matches. Logistic regression was used to assign a weight to this combined keyword matching score, and not to the score for each keyword. On the other hand, a weight was assigned to just the one causal relation "automation->*". It may be that if logistic regression were used to determine an appropriate weight to use for the single keyword "automation", the same or bigger retrieval improvement would have been obtained, and "automation->*" matches would not have contributed anything to improving the retrieval results. The question, then, is:

> **Supplementary question 2**:
> For the queries that did benefit from causal relation matching in the *routing queries* experiment, will the use of causal relation matching still yield an improvement in retrieval results if regression is used to determine the weights not only for the individual causal relation matches but also for the individual keyword matches?

We want to know whether causal relation matching can contribute anything to retrieval effectivenes over and above any contribution by keyword matching.

The analysis reported in this section was designed to obtain indirect evidence that the use of causal relation matching is likely to improve retrieval results over using keyword matching alone. The approach taken was to compare certain conditional probabilities. For example, to show that "automation->*" matches can be used to improve retrieval compared with using just the keyword *automation*, I obtain the following two conditional probabilities:

(1) Prob (REL | "automation" occurs in the document)
The probability that a document is relevant given that the word "automation" occurs in the document.

(2) Prob (REL | "automation->*" occurs in the document)
The probability that a document is relevant given that the relation "automation->" occurs in the document.*

If the conditional probability (2) is greater than the conditional probability (1), then a document containing "automation->*" is more likely to be relevant than a document containing "automation" but not "automation->*". This indicates that the retrieval result is likely to improve if the documents containing the "automation->*" relation are given a higher retrieval score and are ranked above the documents containing the keyword *automation* but not the relation.

So, if probability (2) is greater than probability (1), it suggests that causal relation matching can be used to improve retrieval results over keyword matching. The greater the value of *Prob (2) - Prob (1)*, the bigger the improvement is likely to be. Given a particular positive value of *Prob (2) - Prob (1)*, the actual amount of improvement obtained will depend partly on the following factors:

- the proportion of documents with the causal relation match. The greater the proportion of documents with the causal relation match, the greater the retrieval improvement is likely to be.
- the number of keywords in the query statement and the proportion of documents with high keyword matching scores. Causal relation matching will probably be more useful for those documents containing a few query keywords than for those documents containing many query keywords. Consider the following probabilities:
  (3) Prob (REL | "automation" & "cost" & "decrease" occur in the document)
  (4) Prob (REL | "automation->*" & "cost" & "decrease" occur in the document)

Prob (4) - Prob (3) will probably have a smaller value than *Prob (2) - Prob (1)*.
Even though the effectiveness of causal relation matching is affected by these factors, I think that the value of *Prob (2) - Prob (1)* still provides some indication of how useful causal relation matching is likely to be. In fact, I shall show later that a good predictor of the percentage improvement in retrieval result is the percentage improvement in the conditional probability weighted by the following two factors:

1. the proportion of documents in the test database satisfying the condition used in *Prob (2)* (i.e. the proportion of documents containing the relation "automation->*"), and
2. the proportion of documents in the test database satisfying the condition used in *Prob (1)* (i.e. the proportion of documents containing the keyword "automation").

When is word proximity matching likely to help?

(1) Prob (REL | word1 & word2 occurs in document)
(2) Prob (REL | word1 & word2 co-occurs in sentence)

Word proximity matching is likely to help if (2) > (1).

When is "word1->word2" matching likely to help?

(3) Prob (REL | word1->word2 identified by system in document)
(4) Prob (REL | word1->word2 identified manually in document)

"word1->word2" matching is likely to help if (4) > (2).

When is "word1->*" matching likely to help?

CASE 1: Query does not specify the effect. Only the cause is specified in the query.

(5) Prob (REL | word1 occurs in document)
(6) Prob (REL | word1->* identified by system in document)
(7) Prob (REL | word1->* identified manually in document)

"word1->*" matching is likely to help if (7) > (5).

CASE 2: Query specifies both the cause and the effect

(5a) Prob (REL | word2 not in document, word1 occurs in document)
(6a) Prob (REL | word2 not in document, word1->* identified by system in document)
(7a) Prob (REL | word2 not in document, word1->* identified manually in document)

(5b) Same as (1)
(6b) Prob (REL | word2 occurs in document, word1->* identified by system in document)
(7b) Prob (REL | word2 occurs in document, word1->* identified manually in document)

"word1->word2" matching is likely to help if (7a) > (5a) or (7b) > (5b).

(Continued in the next table)

Figure 9. Description of the conditional probabilities obtained for each query

(Continued from the last table)

When is "*->word2" matching likely to help

CASE 1: Query does not specify the cause. Only the effect is specified in the query.

(8) Prob (REL | word2 occurs in document)
(9) Prob (REL | *->word2 identified by system in document)
(10) Prob (REL | *->word2 identified manually in document)

"*->word2" matching is likely to help if (10) > (8).

CASE 2: Query specifies both the cause and the effect

(8a) Prob (REL | word1 not in document, word2 occurs in document)
(9a) Prob (REL | word1 not in document, *->word2 identified by system in document)
(10a) Prob (REL | word1 not in document, *->word2 identified manually in document)

(8b) Same as (1)
(9b) Prob (REL | word1 occurs in document, *->word2 identified by system in document)
(10b) Prob (REL | word1 occurs in document, *->word2 identified manually in document)

"*->word2" matching is likely to help if (10a) > (8a) or (10b) > (8b).

Figure 10. Description of the conditional probabilities obtained for each query (continued)

For the purpose of the analysis, I considered three types of causal relation matching: 1. "word->word" matching, 2. "word->*" matching and 3. "*->word" matching. 9 and 9 describe the conditional probabilities that were obtained for estimating the usefulness of each type of matching.

For each type of relation matching, two conditional probabilities were obtained -- one probability for the case where the causal relations are identified by the computer program and a second probability for the case where the causal relations are identified manually. For example, in 9, *Prob (4)* is the probability that a document is relevant given that it contains the relation "word1->word2" *as identified by a human being*. *Prob (3)* is the probability that a document is relevant given that my computer program finds the relation in the document. The comparisons of interest are *Prob (4)* versus *Prob (2)*, *Prob (3)* versus *Prob (2)*, and *Prob (4)* versus *Prob (3)*. If *Prob (3)* is greater than *Prob (2)* (the probability of relevance given that there is a word proximity match), it indicates that my retrieval system should have obtained a retrieval improvement with "word1->word2" relation matching in the experiments. If *Prob (4)* is greater than *Prob (2)*, it indicates that "word1->word2" matching is likely to yield a retrieval improvement *if the relation is identified accurately* (i.e. the system performs at the level of a human being). If *Prob (4)* is greater than both *Prob (3)* and *Prob (2)*, it suggests that a bigger retrieval improvement can be obtained with "word1->word2" relation matching by improving the accuracy of the automatic identification of causal relations.

Prob (5) to *Prob (7b)* (9) were obtained to investigate the usefulness of "word->*" matching. For queries that specify only what the *cause* must be but not what the *effect* must be, probabilities (5), (6) and (7) were used for the analysis. For queries that specify both the cause and the effect, probabilities (5a), (6a) and (7a) as well as (5b), (6b) and (7b) were obtained so that a more detailed analysis could be done.

The probabilities obtained to investigate "*->word" matching are similar to those obtained for "word->*" matching, and are given in 9.

Quite often, the cause or effect is represented by more than one word in the text. In such cases, the important words were selected from the *cause* phrase and the *effect* phrase. Thus, "word1" in 9 and 9 can refer to the set of important keywords taken from the cause phrase, and "word2" can refer to the set of important keywords from the effect phrase.

The manual identification of causal relations was done by one of the judges (Judge A) who provided judgments for the evaluation of the computer program reported in Chapter 4. Obviously, the judge could not read all the documents in the test databases. The general guideline followed was to read enough documents to obtain about 30 documents that satisfy the condition for each of the conditional probabilities (4), (7), and (10).

There are two differences between the procedure used in this supplementary analysis and the procedure used in the retrieval experiments reported earlier:

1. For the supplementary analysis, the human judge was instructed to also identify causal relations in which one member of the relation (the cause or the effect) was not specified in the sentence. In contrast, the computer program was designed to look for causal relations in which both members of the relation (the cause and the effect) were specified in the sentence or in adjacent sentences.
2. A stronger type of stemming was used for the analysis. In the retrieval experiments, a "weak" stemmer was used to convert each word in the text to an entry word in the *Longman Dictionary*. As a result, words like *examination* and *examine* were not conflated since each has a separate entry in the *Longman Dictionary*. (In the experiments, I used Roget category codes to conflate such words as *examination* and *examine*. However, Roget category codes were not found to be helpful in causal relation matching.) For this analysis, each query word was stemmed by truncating it at an appropriate place and the database was searched using the truncated keyword.

For the analysis, I selected six queries that had a retrieval improvement from causal relation matching in the *routing queries* experiment. (These were the queries that had the biggest improvement from causal relation matching according to the *11-point average*.) The conditional probabilities for the six queries are given in 23. For the second part of the analysis, I randomly selected ten queries that did not have better retrieval results with causal relation matching. The conditional probabilities for these ten queries are given in 24 and 25. In all three tables, the figures in parenthesis indicate the proportion of documents in the test database that satisfy the condition used in the conditional probability.

How good are the conditional probabilities in predicting the improvement in retrieval results? We can find out by comparing the magnitude of *Prob (2) - Prob (1)* (as given in 23, 24 and 25) with the actual improvement in retrieval results from word proximity matching in the *routing queries* experiment. *Prob (2)* is the probability that a document is relevant if *word1* and *word2* occurs within the same sentence in the document. *Prob (1)* is the probability of relevance if the document contains *word1* and *word2* (not necessarily within the same sentence). So, the percentage improvement in *Prob (2)* compared with *Prob (1)* indicates the percentage improvement in retrieval results that can be obtained using word proximity matching. In 23, 24 and 25, only 12 queries have values for *Prob (2)* and *Prob (1)*. For each of these queries, I calculated the *weighted* percentage improvement in probability using the formula given in 25. I then calculated the Pearson *r* (product-moment correlation) between the weighted percentage improvement in probability and the percentage improvement in retrieval results. The correlation with the different retrieval measures are given in 25. As shown in the table, there is a strong correlation (0.75) with the retrieval improvement as measured using the *3-point average*.

Let us now examine the probabilities given in 23. These are for six queries that obtained a retrieval improvement with causal relation matching. Earlier in Section 5.4.2.1, I have listed in 17 which types of causal relation matching were found by logistic regression to be useful for which queries. In particular, the types of causal relation matching that were found to be helpful for queries 75, 120, 7, 17, 26 and 146 were as follows:

Query No.	Prob (1)	Prob (2)	Prob (3)	Prob (4)		Prob (5)	Prob (6)	Prob (7)	Prob (8)	Prob (9)	Prob (10)
75	0.35 (0.21)	0.25† (0.02)	0.67† (0.02)	0.00† (0.01)	a:	0.20 (0.24)	0.53 (0.08)	0.42 (0.10)	0.00 (0.21)	0.04 (0.14)	0.00 (0.08)
					b:	0.35 (0.21)	0.53 (0.08)	0.45 (0.10)	0.35 (0.21)	0.47 (0.18)	0.39 (0.12)
120	0.15 (0.43)	0.25 (0.10)	0.44† (0.02)	0.58 (0.02)	a:	0.11 (0.15)	0.09 (0.10)	0.33 (0.15)	0.00 (0.38)	0.00 (0.28)	0.09 (0.37)
					b:	0.15 (0.43)	0.19 (0.21)	0.16 (0.23)	0.15 (0.43)	0.20 (0.18)	0.50† (0.02)
7									0.23 (0.64)	0.28 (0.44)	0.47 (0.57)
17									0.86 (0.09)	0.91 (0.03)	0.94 (0.07)
26									0.70 (0.17)	0.74 (0.09)	0.74 (0.09)
146	0.36 (0.60)	0.50 (0.42)	0.50 (0.06)	0.53 (0.06)	a:	0.00 (0.30)	0.00 (0.21)	0.00 (0.42)	0.00‘ (0.01)	0.00† (0.01)	- (0)
					b:	0.36 (0.60)	0.38 (0.37)	0.18 (0.23)	0.36 (0.60)	0.31 (0.34)	0.34 (0.39)

Notes:

1. The figures in parenthesis indicate the proportion of documents in the test database that satisfy the condition used in the conditional probability. For Prob (4), (7) and (10), the proportion of documents may be an estimate based on a sample.
2. Some cells for queries 7, 17 and 26 are empty because the *cause* is not specified in these queries and so "word1->word2" and "word1->*" are not applicable for these queries.
3. † indicates that the probability is estimated from a sample of fewer than 10 documents.

Table 23. Conditional probabilities of relevance for the different types of causal relation matches (for the queries that *did* obtain a retrieval improvement in the routing queries experiment)

Query No.	**Prob (1)**	**Prob (2)**	**Prob (3)**	**Prob (4)**		**Prob (5)**	**Prob (6)**	**Prob (7)**	**Prob (8)**	**Prob (9)**	**Prob (10)**
30	0.54 (0.29)	0.66 (0.09)	1.00† (0.01)	0.86† (0.02)	a:	0.40 (0.05)	0.50† (0.02)	0.50† (0.02)	0.02 (0.60)	0.02 (0.33)	0.00 (0.38)
					b:	0.54 (0.29)	0.74 (0.11)	0.65 (0.12)	0.54 (0.29)	0.53 (0.17)	0.60 (0.13)
38	0.45 (0.21)	0.89† (0.04)	-† (0)	-† (0)	a:	0.29† (0.03)	0.00† (0.00)	-† (0)	0.24 (0.70)	0.25 (0.52)	0.35 (0.42)
					b:	0.45 (0.21)	0.67† (0.04)	0.50 (0.05)	0.45 (0.21)	0.52 (0.12)	0.50 (0.22)
60	0.29 (0.48)	0.35 (0.34)	0.36 (0.07)	0.38 (0.16)	a:	0.00† (0.02)	-† (0)	-† (0)	0.00 (0.38)	0.00 (0.23)	0.00 (0.26)
					b:	0.29 (0.48)	0.32 (0.13)	0.44 (0.28)	0.29 (0.48)	0.37 (0.38)	0.29 (0.38)
70	0.65 (0.05)	0.89† (0.03)	-† (0)	1.00† (0.00)	a:	0.00† (0.02)	0.00† (0.00)	-† (0)	0.00 (0.68)	0.00 (0.32)	0.00 (0.32)
					b:	0.65 (0.05)	0.83† (0.02)	0.75† (0.01)	0.65 (0.05)	0.64 (0.03)	0.83† (0.07)
72	0.15 (0.37)	0.21 (0.10)	0.14† (0.02)	0.00† (0.01)	a:	0.16 (0.19)	0.33 (0.04)	0.00 (0.05)	0.00 (0.21)	0.00 (0.08)	0.00† (0.05)
					b:	0.15 (0.37)	0.18 (0.15)	0.16 (0.17)	0.15 (0.37)	0.16 (0.19)	0.12 (0.22)

Table 24. Conditional probabilities of relevance for the different types of causal relation matches (for the queries that *did not* obtain a retrieval improvement in the routing queries experiment)

(Continued from the last table)

Query No.	Prob (1)	Prob (2)	Prob (3)	Prob (4)		Prob (5)	Prob (6)	Prob (7)	Prob (8)	Prob (9)	Prob (10)
77a						0.08 (0.31)	0.13 (0.19)	0.12 (0.19)			
77b									0.08 (0.31)	0.14 (0.18)	0.10 (0.22)
104	0.21 (0.10)	0.10 (0.03)	-† (0)	0.00 (0.01)	a: b:	0.20† (0.01) 0.21 (0.10)	-† (0) 0.29† (0.02)	1.00† (0.00) 0.25 (0.03)	0.06 (0.61) 0.21 (0.10)	0.08 (0.22) 0.22 (0.06)	0.10 (0.23) 0.17 (0.13)
115	0.52 (0.46)	0.59 (0.32)	0.33† (0.04)	0.39 (0.11)	a: b:	0.00† (0.00) 0.52 (0.46)	-† (0) 0.42 (0.12)	-† (0) 0.39 (0.15)	0.35 (0.48) 0.52 (0.46)	0.40 (0.39) 0.55 (0.41)	0.50 (0.46) 0.62 (0.40)
131	0.05 (0.72)	0.05 (0.58)	0.05 (0.11)	0.00 (0.07)	a: b:	0.00 (0.06) 0.05 (0.72)	0.00 (0.02) 0.02 (0.27)	0.00† (0.02) 0.00 (0.16)	0.00 (0.13) 0.05 (0.72)	0.00 (0.06) 0.04 (0.40)	0.00† (0.04) 0.00 (0.26)
140	0.02 (0.61)	0.02 (0.39)	0.00 (0.05)	0.00 (0.07)	a: b:	0.00 (0.04) 0.02 (0.61)	0.00† (0.01) 0.04 (0.25)	0.00† (0.02) 0.06 (0.31)	0.01 (0.21) 0.02 (0.61)	0.01 (0.18) 0.00 (0.45)	0.08 (0.27) 0.00 (0.49)

Note: Query 77 contains two causal relations. A separate analysis was done for each of them, labeled 77a and 77b.

Table 25. Conditional probabilities of relevance for the different types of causal relation matches (for the queries that *did not* obtain a retrieval improvement in the routing queries experiment) (continued)

	Percentage improvement in			
	11-pt average	**3-pt average**	**normalized recall**	**normalized precision**
_ Prob	0.65*	0.75*	0.19	0.52

Notes

1. _ Prob is the weighted percentage improvement in Prob (2) compared with Prob (1), and is calculated using the formula:

$$\frac{\text{Prob (2) - Prob (1)}}{\text{Prob (1)}} \;*\; \text{size (1)} \;*\; \text{size (2)}$$

where

size (1) is the proportion of documents in the test database satisfying the condition used in Prob (1), and

size (2) is the proportion of document in the test database satisfying the condition used in Prob (2).

(*Size (1)* and *size (2)* are given within parentheses in 23, 24 and 25.)

2. * indicates that the correlation is significant at the 0.05 level (2-tailed test).

Table 26. Correlations between the improvement in the probability of relevance and the improvement in the retrieval results from term proximity matching

Query	Type of causal relation matching that was helpful
75	"word->*" and "*->word pair"
120	"word pair->*"
7	"*->word" and "*->word pair"
17	"*->word"
26	"*->word"
146	"*->word"

If we subsume "word pair->*" matching under "word->*" matching, and "*->word pair" matching under "*->word" matching, we find that the types of causal relation matching selected by the logistic regression generally agree with the conditional probabilities given in 23. For example, for Query 75, we see in 23 that

Prob (6a) >> Prob (5a)
Prob (6b) > Prob (5b)
Prob (9a) > Prob (8a)
Prob (9b) > Prob (8b)

The proportion of documents involved in each of the conditional probabilities is also high. It is no wonder that Query 75 had the biggest percentage improvement in retrieval results from causal relation matching. We also see in 23 that *Prob (3)* is much greater than *Prob (2)* for Query 75 suggesting that the query should also have benefited from "word->word" matching. However, the proportion of documents involved in *Prob (3)* and *Prob (2)* is small (0.02), and so the effect is very small.[46]

Query 146 is unusual in that "*->word" matching obtained a negative coefficient in the logistic regression, indicating that "*->word" matching helped to identify non-relevant documents. This is correctly reflected in the conditional probabilities: *Prob (9b) < Prob (8b)*.

In answer to *supplementary question 2*, the results indicate that even if regression is used to determine the weights for individual keyword matches, the use of causal relation matching is still likely to yield a retrieval improvement -- at least for those queries that had the biggest retrieval improvement from causal relation matching in the routing queries experiment. For five of the queries, documents with causal relation matches are more likely to be relevant than documents containing the keyword(s) involved in the causal relation but not the causal relation itself. For Query 146, documents with causal relation matches are less likely to be relevant and this fact was used by the retrieval system to lower the retrieval scores for these documents.[47]

There are two instances in 23 where a relatively big improvement in the conditional probability was not reflected in the regression models:

Query 120: *Prob (9b) > Prob (8b)*
Query 146: *Prob (6b) > Prob (5b)*

I can think of two reasons why the improvement was not reflected in the retrieval results:

[46] It appears that if the weighted percentage improvement in probability is greater than 0.01, the effect is big enough to be reflected in the retrieval results.

[47] Query 146 asks for documents about peace efforts to end the war in Nicaragua. Many of the documents with causal relations are about factors that exacerbated the situation (i.e. *caused* the situation to deteriorate) or about people getting killed (i.e. *caused* to die).

1. the beneficial effect of the causal relation matches was attenuated by keyword matches
2. the weak stemming that was used in the retrieval experiments. A retrieval improvement from causal relation matching might have been obtained had a stronger stemmer been used.

For both Query 120 and 146, the stronger stemming used for this analysis conflated many more words than the weak stemming used in the retrieval experiments. For example, Query 120 has the relation "*->economic". The weak stemmer used in the retrieval experiments did not conflate the words *economic*, *economical*, *economy,* etc. since these have separate entries in the *Longman Dictionary*. On the other hand, in this analysis, *economic* was simply truncated to "econom" thereby conflating all the variants. This fact and the conditional probabilities suggest that causal relation matching would have yielded better retrieval results had a stronger stemmer been used in the experiments.

For two of the queries, the human judge did substantially better than the system:

Query 120:	*Prob (7a) > Prob (6a)* and *Prob (10a) > Prob (9a)*
Query 7:	*Prob (10) > Prob (9)*

This suggests that a bigger retrieval improvement can be expected with more accurate identification of causal relations.

Let us now examine the probabilities given in 23 and 24. The queries listed in these tables did not obtain a retrieval improvement from causal relation matching. The human judge did substantially better than the system for the following five queries:

Query 38:	*Prob (10a) > Prob (9a)*
Query 60:	*Prob (7b) > Prob (6b)*
Query 104:	*Prob (10a) > Prob (9a)*
Query 115:	*Prob (10a) > Prob (9a)* and *Prob (10b) > Prob (9b)*
Query 140:	*Prob (7b) > Prob (6b)* and *Prob (10a) > Prob (9a)*

In answer to the *first supplementary question*, the results suggest that for about half of the queries that did not obtain a retrieval improvement from causal relation matching, an improvement in retrieval results would have been found if the system had been able to identify causal relations accurately. For these queries, one can expect better retrieval results from causal relation matching by improving the accuracy of the automatic identification of causal relations.

There are seven queries for which the conditional probabilities suggest that a retrieval improvement from causal relation matching should have been obtained in the experiments:

Query 30:	*Prob (6b) > Prob (5b)*
Query 60:	*Prob (9b) > Prob (8b)*
Query 72:	*Prob (6a) > Prob (5a)* and *Prob (6b) > Prob (5b)*
Query 77:	*Prob (6) > Prob (5)* and *Prob (9) > Prob (8)*
Query 104:	*Prob (9a) > Prob (8a)*
Query 115:	*Prob (9a) > Prob (8a)* and *Prob (9b) > Prob (8b)*
Query 140:	*Prob (6b) > Prob (5b)*

On examining the query statements, it appears that 3 of the queries (Query 60, 104 and 115) would have benefited from a stronger stemmer.

Lastly, this analysis provides additional evidence that the most useful types of causal relation matching are those involving a wildcard. "Word->word" matching does not help retrieval because it usually involves such a small percentage of documents. If an effective method is used to expand query terms with synonyms and related terms, this may increase the number of "word->word" matches and hence the usefulness of "word->word" matching.

5.6. Summary

The experiments reported in this chapter were designed to address the following research question:

> **Research question 2**
> Can the information obtained by matching causal relations expressed in documents with causal relations expressed in the user's query statement be used to improve retrieval effectiveness over just matching terms without relations?

The experimental results and the supplementary analysis indicate that, for the type of queries used in the study and for the Wall Street Journal full-text database, causal relation matching can be used to improve information retrieval effectiveness if the weights for the different types of causal relation matching are customized for each query -- as in an SDI or *routing queries* situation.

In the *ad hoc queries* experiment, I used 39 queries to determine the best set of weights to use for combining the subscores from the different types of causal relation matching. However, no retrieval improvement was obtained when the set of weights was applied to a different set of 38 queries.

In the SDI or *routing queries* experiment, in which the best set of weights to use for combining the subscores was determined separately for each query using stepwise logistic regression, causal relation matching yielded a small (2.9%) but significant improvement in the retrieval result, as measured by the *3-point recall-precision average* (p<0.05, 1-tailed test). The retrieval precision was better at most of the recall levels. However, the best results were obtained when causal relation matching was combined with word proximity matching. In this case, the retrieval improvement was significant at the α=0.01 level, and the percentage improvement in the *3-point recall-precision average* was about 4.5%. 31% of the 72 queries used in the experiment obtained better retrieval results using the combination of causal relation matching and word proximity matching than using the baseline keyword matching strategy. 24% of the queries obtained a retrieval improvement from causal relation matching in addition to any improvement obtained using word proximity matching.

A supplementary analysis was carried out to provide another source of evidence that causal relation matching can used to improve retrieval effectiveness. Specifically, the additional analysis was carried out to address two questions:

> **Supplementary question 1**
> For the queries that didn't benefit from causal relation matching in the *routing queries* experiment, will causal relation matching yield an improvement in retrieval results if the automatic identification of causal relations in sentences is more accurate?

> **Supplementary question 2**
> For the queries that did benefit from causal relation matching in the *routing queries* experiment, will the use of causal relation matching still yield an improvement in retrieval results if regression is used to determine the weights not only for the individual causal relation matches but also for the individual keyword matches?

The results indicate that more queries are likely to obtain a retrieval improvement with causal relation matching if the accuracy of the automatic identification of causal relations is improved. The conditional probabilities obtained in the analysis indicate that if causal relations are identified manually, half of the queries analyzed are likely to get better retrieval results from causal relation matching.

With regard to the second supplementary question, the conditional probabilities indicate that, at least for the queries that had the biggest improvement from causal relation matching, the retrieval improvement from causal relation matching is likely to remain even if regression is used to determine the weights for individual keyword matches. For each of the queries analyzed (with the exception of Query 146 explained earlier), the probability of a document being relevant is higher if the retrieval system finds a causal relation of interest in the document than if the system finds the term(s) of the causal relation but not the causal relation itself.

The regression models for the different queries and the conditional probabilities obtained in the supplementary analysis reveal that different types of causal relation matching are helpful for different queries. The most useful types of causal relation matching are those where either the *cause* or the *effect* is a wildcard (i.e. either the *cause* or *effect* is not specified and can match with anything). "Word->word" matching where both the cause and the effect have to find a match is less helpful because such matches are relatively rare. In place of "word->word" matching, word proximity matching should be used. In the *routing queries* experiment, causal relation matching didn't yield better retrieval results than word proximity matching. (Combining causal relation matching with word proximity matching did produce better results than using word proximity matching alone, although the improvement was not significant). Term proximity matching can be seen as a type of causal relation matching in which two terms are assumed to have a causal relation between them if they occur within the same sentence.

In Section 5.2, I posed several subsidiary research questions. I now address the questions using the results from the *routing queries* experiment.

One subsidiary question concerns the number of causal relation matches found in the database during retrieval. If no retrieval improvement from causal relation matching was obtained for a particular query, it might be because the system had difficulty identifying the causal relation of interest to the query.

> **Subsidiary question 2.3d**:
> Is the improvement in retrieval results from causal relation matching greater for queries for which there are more causal relation matches found during retrieval than for queries with fewer causal relation matches?

Retrieval improvement was not found to be higher when there were more causal relation matches in the database than when there were fewer matches. In fact, the reverse appeared to be the case. This may be because a causal relation that occurs very often in the database is not useful in distinguishing relevant from non-relevant documents.

Two of the subsidiary questions refer to term proximity matching:

> **Subsidiary question 2.4a**:
> Is the improvement in retrieval results obtained using causal relation matching greater than the improvement obtained using term proximity matching (i.e. matching causally-related terms in the query statement with terms co-occurring within document sentences)?
>
> **Subsidiary question 2.4b**:
> Does the use of causal relation matching *in addition* to term proximity matching yield better retrieval results than using term proximity matching alone?

The retrieval results obtained using causal relation matching were not better than the results from word proximity matching. However, using causal relation matching in addition to word proximity matching yielded better results than using word proximity matching alone, but the improvement was not significant.

One subsidiary research question asks whether a causal relation match should be given a higher weight when a document has a low score from term matches (without considering relations between terms) than when a document already has a high score from term matches:

> **Subsidiary question 2.2**
> Is there an interaction between term matching score and causal relation matching score in determining the relevance of a document?

A *keyword matching*causal relation matching* interaction term appeared in the regression models for about half of the queries that had a retrieval improvement from causal relation matching. However, several types of causal relation matching were used in this study and no one type of *keyword matching*causal relation matching* interaction term predominated over the others.

Three of the subsidiary questions ask whether the magnitude of the retrieval improvement from causal relation matching depends on some characteristic of the query. Three characteristics were suggested:

1. the centrality or importance of the causal relation to the query
2. the strength of the causal association between the two terms in the causal relation
3. the retrieval results from term matching.

The three subsidiary questions are as follows:

> **Subsidiary question 2.3a**
> Is the improvement in retrieval effectiveness from causal relation matching greater when the causal relation is *central* to the user's query than when the causal relation is *peripheral* to the query?
>
> **Subsidiary question 2.3b**
> Is the improvement in retrieval effectiveness from causal relation matching greater when the causal association between the two terms in the relation is weak than when the association is strong?
>
> **Subsidiary question 2.3c**
> Is the improvement in retrieval results from causal relation matching greater for queries that obtain poor retrieval results from term matching than for queries that obtain good results from term matching?

The retrieval improvement from causal relation matching was greater for queries in which the causal relation was central than for queries in which the causal relation was peripheral. However, the retrieval improvement was not significantly different in the two cases.

Subsidiary question 2.3b relates to the causal association strength between two terms. The causal association between two terms is strong if there is usually a causal relation between the two terms when they occur within a document. I expected the retrieval improvement from causal relation matching to be greater when the causal association between the two terms was weak than when the causal association was strong. This was not found to be the case in the *routing queries* experiment. However, it should be pointed out that the causal association strengths used in the analysis were just my own subjective impressions. What the result actually indicates is that it is difficult for people to predict from reading the query statement whether causal relation matching is going to help or not.

With regard to *subsidiary question 2.3c*, the retrieval improvement from causal relation matching was greater for queries that obtained poor results from keyword matching -- especially when the retrieval improvement was expressed as the percentage of the possible improvement.

Three subsidiary questions relate to the use of Roget category codes for retrieval:

> **Subsidiary question 2.1a**
> Does the use of Roget codes *in place of keywords* improve retrieval effectiveness?
>
> **Subsidiary question 2.1b**
> Does the use of Roget codes *in addition to keywords* improve retrieval effectiveness over using keywords alone?
>
> **Subsidiary question 2.1c**
> Does the use of Roget codes *in addition to keywords* as the terms in causal relations improve retrieval effectiveness?

The results from a preliminary experiment (reported in Appendix 6) show that doing Roget category code matching instead of keyword matching did not give better retrieval results. However, using Roget code matching *in addition to* keyword matching did significantly improve retrieval results in the *routing queries* experiment. What was unexpected was that the weight for the Roget code scores turned out to be negative for many of the queries. Roget code matching seems to capture some *dissimilarity* between query and document that is not captured by keyword matching. A possible reason for this is that many of the important terms in the queries either don't have synonyms or are not listed in *Roget's International Thesaurus*. Only the less important and more common words in the queries tend to have Roget codes. So, Roget codes may be a means of indicating the unimportant words! If a document has a high score from Roget code matching, it may be an indication that many of the keyword matches in the document are for high frequency keywords, and that the keyword matching score for the document should be decreased.

Lastly, using Roget codes in addition to keywords as terms in causal relations did not improve retrieval effectiveness over using just keywords in causal relations.

CHAPTER 6
CONCLUSION

6.1. Introduction

In this study, I developed an automatic method for identifying causal relations in Wall Street Journal text. The automatic method was used in my experimental retrieval system to identify causal relations in documents for the purpose of matching them with causal relations that had been manually identified in query statements. Several retrieval experiments were carried out to investigate whether causal relation matching could be used to improve retrieval effectiveness.

6.2. Automatic Identification of Causal Relations in Text

In developing the automatic procedure for identifying causal relations, I have focused on the causal relations that are explicitly indicated in the text using linguistic means. Many causal relations in text are implied. To infer these implied causal relations require extensive world knowledge, which currently have to be hand-coded. It is thus possible to implement knowledge-based inferencing only for a very narrow subject area. It is currently not possible to use knowledge-based inferencing of causal relations in an information retrieval system that caters to a heterogenous user population with a wide range of subject interests. By focusing on linguistic clues of causal relations, I hoped to develop an automatic method for identifying causal relations that was not limited to a narrow domain and that was accurate enough to be useful for information retrieval purposes.

An evaluation indicates that about 68% of the causal relations in Wall Street Journal text that are clearly expressed within a sentence or between adjacent sentences can be correctly identified and extracted using the linguistic patterns developed in this study (this is the *recall* measure). The result is based on the causal relations that were identified in sample sentences by both of the human judges. Of the instances that the computer program identifies as causal relations, about 72% (the precision) can be considered to be correct -- if we give the computer program the benefit of the doubt when the causal relation is not clearly wrong.

Most of the errors made by the computer program are due to

1. the complexity of the sentence structure
2. problem with lexical ambiguity
3. absence of inferencing from world knowledge.

An analysis of the errors indicates that the use of an accurate parser can improve the *recall* from 68% to as high as 83%, and the *precision* from 72% to 82%. Obtaining or developing an accurate parser is probably the most important step to take to improve the effectiveness of the method.

It should be noted that this study has looked only at causal relations within a sentence or across adjacent sentences. To identify causal relations across larger distances require knowledge-based inferencing.

How well will the approach used in this study work for other corpora? I surmise that the approach will work well with databases containing abstracts of journal articles in a particular subject area -- especially abstracts reporting results of empirical research. Causal relations are probably stated explicitly in such abstracts.

The linguistic patterns developed in this study will have to be tested and modified before they can be used with other corpora. Since the linguistic patterns are based on an extensive review of the literature, they probably include most of the commonly used means of indicating cause and effect. However, each subject area will have its preferred means of indicating cause and effect, and these may not be adequately handled by the linguistic patterns developed in this study.

The linguistic patterns developed in this study are listed in Appendix 4, and the list of causal verbs used are given in Appendix 3. It is hoped that other researchers will find these useful.

6.3. Use of Causal Relation Matching to Improve Retrieval Effectiveness

Does causal relation matching help to improve information retrieval effectiveness? The retrieval results indicate that, for Wall Street Journal text and the kind of queries in the TREC test collection, causal relation matching can yield a small retrieval improvement if the weights for the different types of matching are customized for each query -- as in an SDI or routing queries situation. The best results are obtained when causal relation matching is combined with word proximity matching. Using this strategy, an improvement of 4.5% in the *3-point recall-precision average* was obtained in the *routing queries* experiment. 31% of the 72 queries used in the experiment obtained a retrieval improvement. 24% of the queries obtained a retrieval improvement from causal relation matching in addition to any improvement obtained using word proximity matching. The small size of the retrieval improvement is not unexpected considering that there is usually more than one relation in a query statement. As we incorporate more types of relation matching in the information retrieval strategy, more substantial improvements can be expected.

Perhaps the most important insight obtained in this study is that relation matching where one member (i.e. term) of the relation is a wildcard is especially helpful. The most useful types of causal relation matching were found to be those where either the *cause* or the *effect* was not specified and could match with anything. Wildcard matching is helpful in the following cases:

1. when the document uses a synonym or related term that is not anticipated by the user. Using wildcard matching allows the retrieval system to register a partial relation match when there is no match for one member of the relation.
2. when one member of the relation is specified in a different sentence in the document, as in the following two examples:

   (1) The policeman surprised a burglar.
   In the ensuing struggle, he *killed* the burglar.

   (2) The policeman surprised a burglar.
   The burglar *was killed* in the ensuing struggle.

   In both examples, there is a causal connection between the policeman and the burglar's death. In example (1), an anaphor is used in the second sentence to refer to the policeman in the first sentence. In example (2), the policeman is not referred to in the second sentence but is implied by the context.

3. when one member of the relation is not specified in the query but is expected to be supplied by the relevant documents, as in this example query:

   I want documents that describe the consequences of the Gulf War.

   In this case, the user is not able to specify his information request completely. In the terminology of Belkin, Oddy and Brooks (1982a and 1982b), the user has an *Anomalous State* *of Knowledge* (ASK). Belkin *et al.* said that the goal of information retrieval is to resolve anomalies in a person's state of knowledge by retrieving documents whose content will remove the anomaly. For the type of *ASK* exemplified in the example query, the anomaly can be represented as a wildcard in a causal relation:

Gulf war -> *

The retrieval system will attempt to retrieve documents containing a relation that will instantiate the wildcard and so remove the anomaly.

"Word->word" matching where both the cause and the effect have to find a match is less helpful because such matches are relatively rare. In place of "word->word" matching, word proximity matching should be used. Word proximity matching was found to give significantly better retrieval results than the baseline keyword matching method. Word proximity matching can be seen as a type of causal relation matching in which two terms are assumed to have a causal relation between them if they occur within the same sentence.

One feature of this study is the use of Roget category codes for term conflation. Replacing keywords with Roget category codes was not found to improve the retrieval results. However, using Roget code matching *in addition to* keyword matching did significantly improve retrieval results in the *routing queries* experiment. What was unexpected was that the weight for the Roget code scores turned out to be negative for many of the queries. Roget code matching seemed to capture some *dissimilarity* between query and document that was not captured by keyword matching. A possible reason for this is that many of the important terms in the queries either don't have synonyms or are not listed in *Roget's International Thesaurus*. This limits the usefulness of using a general purpose thesaurus like Roget's. It may be the reason why using Roget category codes as terms in causal relations did not yield a retrieval improvement with causal relation matching.

Can the beneficial effect of causal relation matching on retrieval effectiveness be enhanced in some way? There are two factors that may increase the retrieval improvement from causal relation matching:

1. more accurate identification of causal relations
2. more effective query expansion.

The supplementary analysis using manually identified causal relations indicates that improving the accuracy of the automatic identification of causal relations is likely to yield bigger retrieval improvement and also result in more queries benefiting from causal relation matching.

I surmise that using an effective method of expanding query terms with synonyms and related terms can enhance the usefulness of causal relation matching. Expanding the terms in the causal relations with additional alternative terms increases the chances of a causal relation match. Query expansion is primarily a recall enhancement device. Adding more terms to the query will help to retrieve more of the relevant documents, but will also cause more of the non-relevant documents to be retrieved as well. Relation matching can reduce the number of non-relevant documents retrieved by the additional query terms.

Are the results obtained in this study applicable to other databases and user populations? I expect the results to hold for other newspaper texts. Whether the same results would have been obtained with other types of text and queries is a question for future research. Obvious factors that affect the generalizability of the results are:

1. the type of database records (e.g., full-text documents or document surrogates)
2. the type of query (e.g., long or short query statements, and whether relations are important to

the user)
3. the type of documents that the database indexes (e.g., newspaper articles, journal articles, books or patents)
4. the subject area.

Relation matching is more likely to be useful with full-text databases than with databases of document surrogates (e.g., abstracts). The reasons for this prediction are:

1. Full-text documents contain many more terms than document surrogates, and so, term matching is more likely to produce false drops (i.e. non-relevant documents that are erroneously retrieved). Relation matching will help to increase the precision of retrieval.
2. With a full-text database, there is a greater likelihood of relation matches. A relation is more likely to be expressed several times in different ways in the full text of a document than in the document's abstract. The retrieval system is, thus, more likely to detect at least one occurrence of the relation in the full-text document, and this means a higher accuracy in identifying documents that contain the relation of interest to the user.
3. A relation that is expressed in a document may not appear in its abstract. With a full-text database, the retrieval system can make use of details of the user's information need (including relations between concepts) in attempting to identify relevant documents. With document surrogates that summarize the content of documents, the retrieval system may have to use broader concepts in the search, and details of the users' information need is less likely to be useful in the search.

With regard to databases of abstracts, relation matching is more likely to be useful with long informative abstracts than with short descriptive abstracts.

The TREC query statements used in this study specify what information a document must contain to be relevant. Relevance judgment was based on whether the document contained the desired information, rather than on the subject of the document as a whole. Most of the TREC query statements contain a few concepts and relations (i.e. they are not one-concept queries). Relation matching is more likely to be useful for long query statements that specify the users' information need in detail than for short queries that specify the general subject of the documents to be retrieved.

Relation matching is also more likely to be useful for queries that place greater emphasis on the relations between concepts. In this study, better results from causal relation matching was obtained for queries in which the causal relation was central than for queries in which it was peripheral, although the difference was not significant.

Whether relation matching will improve retrieval results also depends on how accurately the relation can be identified by the retrieval system. Automatic detection of relations is likely to be more accurate if:

1. the relations of interest to the user population tend to be explicitly expressed in the text
2. the sentence structure of the text tends to be simple.

To what extent the text meets the above two conditions depends partly on the type of document and the subject area. I surmise that causal relations are more likely to be explicitly expressed in reports of empirical research and in the abstracts of such reports than in the newspaper text used in this study. I further surmise that such documents generally contain simpler sentence structures than do *Wall Street Journal* documents. A high proportion of the sentences in *Wall Street Journal* are long and complex. The writers seem to be trying to express as much information as possible in as few words as possible.

The usefulness of relation matching also depends on what I termed in Chapter 1 "relational ambiguity." If there is usually one particular relation between term A and term B whenever they occur within a document, then relation matching between term A and term B will not improve retrieval. The relation between term A and term B can be assumed simply from the co-occurrence of the two terms in a document. It is only when a variety of different relations are possible between two terms that relation matching between the two terms can improve retrieval. Relational ambiguity between query terms in documents may vary with the subject area. A study of "relational ambiguity" in the literature of different subject areas will throw light on the possible usefulness of relation matching in those subject areas.

6.4. Directions for Future Research

There are several other unanswered questions relating to the use of causal relations in information retrieval. I shall highlight only a few of them.

Identifying causal relations in text is computationally expensive. Most of the complexity is in identifying which part of the sentence is the cause and which the effect. This study made use of a part-of-speech tagger and a phrase bracketer, but not a full-fledged parser. So, the linguistic patterns developed in this study have to perform some of the functions of a parser in order to identify accurately which part of the sentence represents the cause and which the effect. Detecting causal relations will be much simpler if the system does not have to extract the cause and the effect, but only has to recognize that there is a causal relation somewhere in the sentence. In this case, identification of causal relation becomes simply a matter of keyword matching, since the linguistic patterns use mainly keywords to detect causal relations. In other words, to identify the causal relation

word1 -> word2

we can simply search for the co-occurrence of "word1", "word2" and "cause" (with all its "synonyms" and other lexical indications of its presence) within the same sentence. It would be interesting to see if this approach gives better results than the more sophisticated approach attempted in this study.

In this study, I was able to obtain some retrieval improvement in the *routing queries* experiment but not in the *ad hoc queries* experiment. In the *routing queries* experiment, a large number of relevance judgments were used to find out which types of relation matching were useful and how each should be weighted. The question is how to predict which types of relation matching will improve retrieval in an *ad hoc queries* situation where there is little or no relevance feedback. Most information searches are of the ad hoc type, so finding a way of using relation matching without extensive relevance judgments is important.

I view this study as the beginning of a program of research to explore the use of relations in general for improving retrieval effectiveness. Future studies can explore the use of other types of semantic relations. In this study, I have focused on just one relation and have tried to develop an automatic way of identifying the relation accurately. Because I did not limit the study to a narrow subject area, it was not possible to use knowledge-based inferencing. It is difficult to identify semantic relations accurately without knowledge-based inferencing.

I think that many useful insights can be obtained by focusing on a narrow subject area and doing a case study of a small number of queries. The study can attempt to handle all the relations expressed in the queries. Since the study will be limited to a narrow domain, knowledge-based identification of the relations can be implemented, and the usefulness of relation matching for information retrieval can be investigated more thoroughly.

One of my motivations for focusing on the causal relation and for paying close attention to the accurate extraction of cause and effect from text was to eventually develop a computer program for extracting useful causal knowledge from a textual database. I have extracted cause-effect information from about five years of Wall Street Journal articles (about 175,000 articles) but have not investigated how good this collection of cause and effect information is nor what use it can be put to.

One possible use of causal knowledge in information retrieval is in query expansion. Causal relation can be used as a paradigmatic relation to expand query terms with causally related terms. In particular, a query term can be expanded with

1. other terms that tend to cause the query term. For example, if a query statement contains the term "lung cancer", it can be expanded with "cigarette smoking", "air pollution" as well as the terms for other factors that are likely to cause lung cancer.
2. other terms that tend to be caused by the query term. So, the query term "cigarette smoking" can be expanded with "lung cancer", "asthma", "respiratory tract diseases", and the names of other diseases that are likely to be caused by cigarette smoking.
3. other terms that share the same cause (i.e. can be caused by the same factor). The query term "lung cancer" can be expanded with "respiratory tract diseases" because they are both likely to be caused by cigarette smoking.
4. other terms that have the same effect (i.e. can cause the same thing). The query term "cigarette smoking" can be expanded with "air pollution" because they both can cause lung cancer.

Causal knowledge extracted from textbooks and full-text databases are potentially useful in knowledge-based systems (Chorafas, 1990; Kaplan & Berry-Rogghe, 1991) and in intelligent decision support systems (Venkatesh, Myaeng & Khoo, 1994).

Besides extracting causal knowledge, a computer program for identifying causal relations can also assist in synthesizing new causal knowledge. If one document states that A causes B, and another document states that B causes C, the system can be programmed to chain the two causal relations to suggest the inference that A causes C. Swanson (1986, 1991 and 1990) has been studying the possibility of infering new medical knowledge from the medical literature and has found a few examples of such "undiscovered public knowledge."

APPENDIX 1. INSTRUCTIONS TO JUDGES ON HOW TO IDENTIFY CAUSE AND EFFECT IN SENTENCES

This list contains 600 pairs of sentences. For each pair of sentences, first decide whether it contains any cause-effect relation. The cause-effect relation may be between the two sentences or between two phrases within a sentence. For each cause-effect relation that you find, mark with brackets [^{C} ...] the phrase or sentence representing the cause. Also mark with brackets [^{E} ...] the phrase or sentence representing the effect. For example:

> [^{E} The car didn't brake in time.]
> This was because [^{C} the road was slippery.]
>
> The doctor advised the patient to stop smoking.
> He said that [^{C} cigarette smoking] can cause [^{E} lung cancer.]

<u>There may be more than one cause-effect relation</u>, for example:

> [^{E} The car didn't brake in time] because [^{C} the road was slippery] due to the heavy rain.
> [^{E}] [^{C}]

You should indicate only the cause and effect that is <u>explicitly expressed</u> in the sentence. For example, in the sentence

> John had a car accident and was taken to the hospital in an ambulance.

there is an implied cause-effect relation between John's accident and his trip to the hospital. The cause and effect is only implied and not explicitly indicated in the sentence. I am not interested in implied cause and effect.

There are many ways in which a cause-effect relation can be expressed in a sentence. The following are some of the ways. Note that these may not be the only ways of expressing cause and effect.

1. Causal link

A causal link is one or more words that link two phrases or clauses, and indicates a cause-effect relation between them. The following are examples.

a. *Causal link between two clauses*

> [^{C} There was a lot of snow on the ground.] *For this reason* [^{E} the car failed to brake in time.]
> [^{C} There was a lot of snow on the ground] *with the result that* [^{E} the car failed to brake in time.]
> [^{C} It was snowing heavily,] and *because of the snow* [^{E} the car didn't brake in time.]
> [^{C} There was an unexpected snow storm over the holiday weekend,] *with the following consequences*: [^{E} . . .]

b. *Causal link introducing a prepositional phrase*

> [^{E} The car failed to brake in time] *because of* [^{C} the slippery road.]
> [^{E} The car crash,] *due to* [^{C} slippery road conditions,] could have been avoided had the road been cleared of snow.

c. *Causal link introducing a subordinate clause*

[^{E} The car didn't brake in time] *because* [^{C} the road was slippery.]
[^{C} *Being* wet,] [^{E} the road was slippery.]
[^{C} There was *so* much snow on the road] *that* [^{E} the car couldn't brake in time.]

d. *Causal link introducing the subject or complement of a clause*

[^{E} The car accident] was *due to* [^{C} the slippery road.]
The *reason* [^{E} the car didn't brake in time] was *because* [^{C} the road was slippery.]
[^{C} The heavy rain] has *these effects*: [^{E} . . .]

2. "If ... then ..." sentences

"If ... then ..." sentences often indicate a cause-effect relation:

If [^{C} you work hard,] [^{E} you will pass the test.]
If [^{C} you mow the lawn] then [^{E} I will give you $5.]

Not all *if-then* sentences indicate a cause-effect relation, however:

If you see lightning, you will soon hear thunder.

Although you always hear thunder after you see lightning, one does not cause the other since both thunder and lightning are caused by the same event.

Other examples of *if-then* sentences that don't indicate a cause-effect relation:

If you see the gas station, you have already passed my house.
If you visit St. Louis, be sure to see the arch.
If she's poor, at least she's honest.

3. Causal verbs

There are some transitive verbs that indicate cause and effect. For example, the sentence

John *broke* the vase.

can be paraphrased as

John caused the vase to break.

The cause is *John* and the effect is that *the vase broke*:

[^{C} John] [^{E} broke the vase.]

Other verbs like *hit* do not express cause and effect:

John *hit* the ball with a bat.
John *hit* the vase with a bat.

This sentence does not explicitly say what effect John's hitting had on the ball or vase, so there is no explicit cause-effect relation in this sentence. In the previous sentence involving the verb *break*, the effect on the vase is expressed (i.e. the vase broke). So, there is a cause-effect relation in the previous sentence.

Other examples of causal verbs are:

Verb	Paraphrase
move	to cause to move
melt	to cause to melt
kill	to cause to die
destroy	to cause to be destroyed
convince	to cause to believe
raise	to cause to rise
anger	to cause to be angry
shelve	to cause to be on the shelf

Example sentences:

[^{C} John] [^{E} moved the box.] (i.e. John caused the box to move.)

[^{C} The heat from the fireplace] [^{E} is melting the ice-cream.]

[^{C} The earthquake] [^{E} destroyed the city] , killing thousands of people.
[^{C}] [^{E}]
(Note that there are two cause-effect relations in this sentence.)

[^{C} Watching her neighborhood kids get shot] [^{E} convinced Jane that firearms should be banned.]

[^{C} The government] [^{E} will raise taxes] next year.

[^{C} His action] [^{E} angered many people in the community.]

[^{C} Bob] [^{E} has shelved the book]

There are three special groups of transitive verbs that primarily mean "to cause":

- coercive causatives: e.g. coerce, force, compel
- neutral causatives: e.g. cause, result in, result from, lead to
- permissive causatives: e.g. allow, permit, let

There is also a group of transitive verbs that mean "to cause to NOT happen", for example:
prevent, avert, forestall, foil
Example: [^{C} Prompt response from the police] [^{E} averted disaster.]

The following transitive verbs are NOT causal verbs:
measure, cite, risk, see, kick, eat, kiss

Example sentences containing the above verbs that are NOT causal:
The workman *measured* the room.
He *cited* three studies that disproved the theory.

He *risked* a lot of money in the venture.
I *saw* him in the store.
Bart *kicked* the referee.
Homer *ate* an apple.
Pasquale *kissed* his mother.

4. Resultative constructions

The following sentences indicate a cause-effect relation:

[^{C} Grandpa kissed **[^{E}** my nose] wet. **]** (I.e., Grandpa kissed my nose and that caused my nose to be wet.)
[^{C} The maid wiped **[^{E}** the dishes] dry. **]**
[^{C} The grocer ground **[^{C}** the coffee beans] to a fine powder. **]**
[^{C} They painted **[^{C}** their house] a hideous shade of green. **]**

[^{C} The alarm clock ticked] [^{E} the baby awake.]
[^{C} The dynamite blew] [^{E} the safe open.]

5. Additional instructions relating to tense and modality

Please include cause-effect relations expressed in any of the verb tenses.

Examples:
[^{C} John] [^{E} broke the vase] yesterday.
[^{C} John] [^{E} is always breaking vases.]
[^{C} John] [^{E} has been breaking vases] since he was a child.
I'm sure that [^{C} John] [^{E} will break the vase.]

Also include cause-effect relations that are expressed as *intended* or *possible* cause and effect.

Examples:
[^{C} John] has the intention of [^{E} breaking the vase.]
[^{C} John] tried [^{E} to break the vase.]
Ted told [^{C} John] [^{E} to break the vase.]
[^{C} John] agreed [^{E} to break the vase.]
[^{C} John] [^{E} may break the vase.]
John is so clumsy [^{C} he] [^{E} is sure to break the vase.]

[^{E} The accident] may be the result of [^{C} the slippery road conditions.]

However, exclude cause-effect that is *negated*.

Examples:
It is true that John did not break the vase.
It is not true that John broke the vase.
John denies breaking the vase.
John failed to break the vase.
John refused to break the vase.
John may not have broken the vase.

The plane crash was not due to pilot error.
He failed the exam not because he didn't study hard, but . . .
The cigarette manufacturers claim that cigarette smoking is not a cause of lung cancer.

Summary of the various ways of indicating cause and effect

1. Causal links

... for this reason ...
... with the result that ...
... because of ...
... due to ...
... because ...
... The reason was ...
... has these effects: ...

2. Conditionals

If ... then ...

3. Causal verbs

Verb	Paraphrase
break	to cause to break
move	to cause to move
melt	to cause to melt
kill	to cause to die

4. Resultative constructions

[^{C} The grocer ground **[**^{C} the coffee beans] to a fine powder. **]**
[^{C} The dynamite blew] [^{E} the safe open.]

APPENDIX 2. LIST OF PART-OF-SPEECH TAGS AND PHRASE MARKERS USED IN TEXT PROCESSING.

130

Contents

Appendix 2

A. Part-of-Speech Tags Used by the POST Tagger (from BBN Systems and Technologies)

Note: The POST tagger adds a "|" character followed by part-of-speech tag to the end of each word.

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential "there"
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	"to"
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb
"	Simple double quote
$	Dollar sign
#	Pound sign
`	Left single quote
'	Right single quote
(	Left parenthesis (round, square, curly or angle bracket)
)	Right parenthesis (round, square, curly or angle bracket)
,	Comma
.	Sentence final punctuation
:	Mid-sentence punctuation

B. Phrase Markers Used by the Phrase Boundary Bracketer Developed in the DR-LINK Project

C	Clause
N	Noun phrase
M	Conjunction of two noun phrases
P	Prepositional phrase
D	Past participle phrase (non-finite clause beginning with past participle verb)
G	Present participle phrase (non-finite clause beginning with a present participle verb)
W	Main verb of a clause, a past participle phrase or a present participle phrase

C. Sample Text Tagged with Part-of-Speech Labels and Phrase Markers

[C [N The|DT West_German|NP central|NN bank|NN]N [W said|VBD]W [C [N the|DT nation|NN 's|POS economy|NN]N [W showed|VBD]W [N no|DT real|JJ growth|NN]N [P in|IN [N the|DT fourth|JJ quarter|NN]N]P [P from|IN [N the|DT previous|JJ three|CD months|NNS]N]P ,|, largely|RB [P because|RB of|IN [G worsening|VBG [N trade|NN conditions|NNS]N]G]P]C]C .|.

[C [P By|IN [N contrast|NN]N]P ,|, [N gross|JJ national|JJ product|NN]N [P in|IN [N the|DT third|JJ quarter|NN]N]P [W grew|VBD]W [N 1%|CD]N [P from|IN [N the|DT second|JJ period|NN]N]P]C .|.

[C [P For|IN [N the|DT full|JJ year|NN]N]P ,|, [N the|DT central|JJ bank|NN]N [T estimated|VBN]T that|IN [M [N the|DT nation|NN 's|POS total|JJ output|NN of|IN goods|NNS]N and|CC [N services|NNS]N]M ,|, after|IN [N adjustment|NN]N [P for|IN [M [N prices|NNS]N and|CC [N seasonal|JJ factors|NNS]N]M]P ,|, [W grew|VBD]W [N 2.5%|CD]N]C .|.

[C Separately|RB ,|, [N a|DT confidential|JJ report|NN]N [T prepared|VBN [P for|IN [N Economics_Minister_Martin_Bangemann|NP]N]P]T [W indicated|VBD]W [C that|IN [N the|DT government|NN 's|POS prediction|NN]N [P for|IN [N 2.5%|CD growth|NN]N]P [P in|IN [N its|PP$ 1987|CD economic|JJ report|NN]N]P [W was|VBD]W too|RB optimistic|JJ]C]C .|.

[C [N Most|JJS of|IN West_Germany|NP 's|POS private|JJ economic|JJ institutes|NNS]N [W expect|VBP]W [N growth|NN of|IN 2%|CD]N at|IN best|JJS]C .|.

APPENDIX 3. LIST OF CAUSAL VERBS.

133

Appendix 3

Contents: 133

A. Verbs that mean *to cause something*
B. Verbs that mean *to be caused by something*
C. Verbs that mean *to prevent something from happening*
D. Verbs that mean *to affect something* without specifying in what way

Note:

- Prepositions within brackets that are preceded by the "+" sign indicate the prepositional phrase that is to be used with the verb.
- Prepositions within brackets that are *not* preceded by the "+" sign indicate the particle that is to be used with the verb.
- The phrase (if any) after the ":" sign expresses the effect of the action denoted by the verb

A. Verbs that mean *to cause something*

A1a. Verbs that are primarily causal in meaning, and where the subject of the verb can be an event

assure
bring
bring (+on) : be experienced by
bring (+prep;particle)
bring (about)
bring (down+on)
bring (on)
catalyze : happen or speed up
cause
compel : exist
effect
engender
ensure
ensure (+that)
eventuate in
ferment

generate
ignite
incite
kindle
lead to
occasion
precipitate
present (+to;+with) : have or be experienced
prompt
provoke
rekindle
result in
spark
trigger

A1b. Verbs that are primarily causal in meaning, and where the subject of the verb cannot be an event but has to be a state, an object or an agent.

call (forth) : appear
contrive
enforce
engineer

foment
force (+prep;particle)
get (+adj) : be
get (+v-ed) : be

get (+v-ing) : be
have (+adj) : be
have (+v-ed)
mediate
put (+prep;particle)
render (+adj) : be
stir (up)
strike (+adj) : be
work (+adj;+prep;particle) : be
wreak

A2. Verbs that mean *to force (someone) to (do something)*

banish : leave
blackmail (+into)
bludgeon (+into)
boot (out) : leave
browbeat (+to-v;+into)
can : leave a job
cashier : leave the armed forces
cast (out) : leave
coerce (+to-v;+into)
coerce : into obedience
compel (+to-v)
compel (+to;+into)
condemn (+to)
condemn (+to-v)
conscript : serve in the army, navy or air force
constrain (+to-v)
defrock : leave the priesthood
deport : to leave the country
depose : leave a position of power
detail (+to-v)
disbar : leave the legal profession
dismiss : leave a job
displace : leave a certain place
doom (+to-v)
dragoon (+into)
drive (off) : leave
drum (+out) : leave
eject : leave
evict : leave
exile : leave
expatriate : leave a country
expel : leave
extradite : go back
flood (out) : leave
flush (+prep;particle) : leave
force (+out_of;+from)
force (+to-v;+into)
frogmarch : move forward with arms held firmly behind
high-pressure (+to-v;+into)
intimidate (+into)
kick (out) : leave
make (+v)
march (+to;+into;particle) : go on foot
muzzle : keep silent
obligate (+to-v)
oblige (+to-v)
oust : leave
pension (off) : retire
pressgang (+to-v;+into)
railroad (+into)
repatriate : go back to their own country
rout (out) : leave
run (+out_of) : leave
rusticate : leave
sack : leave a job
sandbag (+into)
shanghai (+to-v;+into)
slap (down) : into silence or inactivity
smoke (out) : come out
spur : go faster
terrorize (+into)
turf (out) : leave

A3. Verbs that mean *to persuade or cause (someone) to (do something)*

awe (+into)
bamboozle (+into) : to
beat (down) : reduce the price
bribe (+into;+to-v)
cajole (+to-v;+into)
coax (+prep;particle)
coax (+to-v;+into)
con (+into)
convert : accept a particular religion or belief
deceive (+into)
decoy (+into) : be in
delude (+into)
determine (+to-v)
discharge : leave
disembark : disembark
dupe (+into)
embolden (+to-v) : to

entice (+to-v;+prep;particle)
fob (+off_with) : accept
fool (+into)
galvanize (+into)
get (+to-v)
give (+to-v)
goad (+into)
goad : do something
have (+v)
hoodwink (+into)
hustle (+to;+into)
impel (+to)
impel (+to-v)
incite (+to-v)
induce (+to-v)
inspire (+to-v)
instigate (+to-v)
inveigle (+into)
jockey (+into;particle)
jolly (+into)
lead (+to-v)
mislead : think or act mistakenly
persuade (+to-v;+into)
precipitate (+into)
reason (+into)
rope (in) : help in an activity
rush (+into) : act quickly
scare (+prep;particle)
seduce (+prep;particle)
seduce : have sex
shame (+into)
soft-soap (+into)
stampede (+into) : to
stimulate (+to-v)
suborn : do wrong
sweet-talk
talk (+into)
touch (+for) : give
trap (+into) : to
trick (+into)
urge (+prep;particle)

A4a. Verbs that mean *to let or allow (someone) to (do something)*

admit : enter
allow (+prep;particle) : come or go
allow (+to-v)
enable (+to-v)
let (+v)
let (+into) : come into
let (down) : go down
let (in) : come in
let (off) : avoid something
let (out) : leave
permit (+to-v)
set (down) : get out
suffer (+to-v)

A4b. Verbs that mean *to let or allow (an event or state) to happen or to continue to happen*, or *to make (an event or state) possible*

allow
enable
permit
tolerate

A5a. Verbs that mean *to cause (an event) to start*

commence
ignite : start to burn
inaugurate
initiate
instigate
kindle : start to burn
light : start to burn
rekindle : start to burn again
set (+v-ing)
start
turn (on) : start flowing

A5b. Verbs that mean *to bring (something) into existence*, or *to produce (something)*

bear
beget
begin
bioengineer
bore
brew
bring (forth)
build
burrow
chisel
churn (out)
cobble
cofound
coin
compile
compose
concoct
constitute
construct
contour
contrive
craft
crank (out)
crayon
create
crochet
dash (off)
dig
draft
draw (up)
empanel
erect
erect : be erected
establish
excavate
fabricate
fashion
fix
forge
form
formulate
found
generate
grind (out)
hew
impose
inspire
institute
invent
knit
knock (together)
knock (up)
lithograph
machine
make
manufacture
mass-produce
mint
mix
mould
nibble (+prep;particle)
originate
pitch : be erected
prefabricate
prepare
print
procreate
produce
put (forth)
put (together)
put (up)
rebuild
reconstruct
recreate
reduplicate
regenerate
remake
rig (up)
rough (out)
rub (+prep;particle)
scare (up)
sculpture
scythe
secrete
set (up) : be erected
spawn
spin (off)
start
synthesize
tailor (+prep;particle)
tear (off)
throw (together)
tool
toss (off)
tunnel
wear (+prep;particle)
weave

A6. Verbs that mean *to cause (an event or state) to continue*, or *to maintain or preserve (something)*

buoy : continue to float
buoy : remain high
continue : continue
continue : start again

hold (+prep;particle) : remain in position
keep (+v-ing) : continue
maintain : remain in good condition
nourish : remain alive
perpetuate : continue to exist
preserve : be preserved
preserve : remain unchanged
propagate : continue to exist or increase by producing descendants
sustain : keep up the strength, spirits or determination
sustain : remain in existence

A7. Verbs that mean *to cause (something) to operate or to become more active*, or *to cause (something) to come back into use or existence*

activate : become active
activate : happen more quickly
actuate : act
arouse : become active
arouse : become awake
assemble : be assembled
awake : become active or conscious
awake : become awake
awaken : become awake
ginger (up) : become more active and effective
install : ready for use
mobilize : be ready to start working
reactivate : become active again
ready : be ready
reanimate : come back to life; have new strength or courage
recharge : be charged up again
resurrect : come back to life
resuscitate : come back to life
revitalize : have new strength or power; come back to life
revive : become conscious or healthy again
revive : come back into use or existence
revivify : have new life or health
rouse : become awake
rouse : become more active or interested
wake : become awake
waken : become awake

A8a. Verbs that mean *to put (something) out of existence*, or *to destroy (something)*

annihilate : be destroyed
assassinate : die
behead : die by removing head
blast (+prep) : be destroyed
blast (particle) : be destroyed
blow (away) : die
break (down) : be destroyed
bump (off) : die
burn (off) : be destroyed
burn : burn
butcher : die
crucify : die
decapitate : die by removing the head
decimate : be destroyed
demolish : be destroyed
destroy : be destroyed
destroy : die
devastate : be destroyed
disband : breakup and separate
disintegrate : be destroyed
dismantle : be dismantled
dispel : disappear
dissolve : end or break up
drown : die
dynamite : be destroyed
electrocute : die
eliminate : die
execute : die
exterminate : die
extinguish : be put out
extirpate : be destroyed
finish (off) : die
garrotte : die
gas : die
guillotine : die by removing head
gun (down) : die
gut : be destroyed
incinerate : be destroyed
kill (off) : die
kill : die
knock (down) : be destroyed
level : be destroyed
liquidate : die
make away with : die
massacre : die
mow (down) : die
murder : die
obliterate : be destroyed

off : die
overlie : die
poison : die or be harmed with poison
pull (down) : be destroyed
put (down) : die
put (out) : be put out
quench : be put out
ravage : be destroyed
raze : be destroyed
root (out) : be destroyed
ruin : be destroyed
ruin : be ruined
sabotage : be destroyed
shoot (down) : be destroyed
slaughter : die
slay : die
snuff : be put out
strangle : die
strike (+prep;particle) : be removed from
strike (down) : die
suffocate : die by suffocating
tear (down) : be destroyed
tear (up) : be destroyed
trash : be destroyed
wreck : be destroyed
zap : be destroyed

A8b. Verbs that mean *to cause (an event or state) to come to an end*, or *to stop (an event or state) that has been happening*, or *to cause (something) to fail*

abate
abolish : end
abort : end
abrogate : end
annul : end
arrest : end
axe : end
balk
beat : fail
break (+of) : be cured of
cancel : end
conclude : come to an end
defeat : fail
demobilize : end
desegregate : end racial segregation
disable : be no longer able to operate
disable : be unable to use his/her body properly
discontinue : end
dish : fail
dismantle : end
do away with : end
eliminate : end
end : end
eradicate : end
flag (down) : stop
freeze : be unable to work
fuse : stop working
halt : halt
incapacitate : be incapacitated
kill : end
liquidate : liquidate
overturn : end
paralyse : stop working
phase (out) : end gradually
prorogue : end
quell : end
rein (back) : stop
repeal : end
rescind : end
revoke : end
sever : end
snuff (out) : end
stall : stall
stamp (out) : end
staunch : stop discharging blood
stem : stop flowing
still : end
stop : end
stub (out) : stop burning
suppress : end
suspend : not take part in a team for a time
suspend : stop
sweep (away) : end
terminate : end
turn (off) : stop flowing
turn (out) : stop working

A8c. Verbs that mean *to cause (something) to have no effect*

deactivate : be inactive or ineffective
decommission : no longer be in service
invalidate : be invalid
negate : have no effect
neutralize : have no effect
nullify : have no effect
nullify : have no legal force
void : have no effect

A9. Verbs that mean *to cause (something) to be performed or to succeed*

bring (+through) : come through successfully
bring (off) : succeed
complete : be complete
consummate : be complete
effectuate : be successfully carried out
execute : be performed
finalize : be complete
implement : be performed
pull (off) : succeed
push (through) : be successful or accepted
railroad : pass or be implemented
rerun : be held again
solemnize : be performed
stage : be performed
stitch (up) : be complete

A10. Verbs that mean *to cause (something) to be removed*, or *to cause (something) to have something removed from it*

bark : have the skin rubbed off
bone : have the bones removed
comb (out) : be removed
core : have the core removed
cross (off;out) : be removed
de-ice : be free of ice
debeak : have the beak removed
debone : have the bones removed
declaw : have the claw removed
defrost : be free of ice; thaw
defrost : be free of steam
defuse : have the fuse removed
dehumidify : have moisture removed
delete : be removed
demagnetize : be free of magnetic qualities
demagnetize : be free of sounds
demist : be free of steam
denude : have the protective covering removed
desalinate : be free of salt
descale : be free of scale
detoxify : have the poison removed
disafforest : have the forest removed
disarm : have the weapons removed
disembowel : have the bowels removed
disentangle : have knots removed
dislodge : be dislodged
dismast : have the masts removed
dismember : have the limbs torn off
divest (+of) : be removed of
edit (out) : be removed
efface : be removed
erase : be removed
eviscerate : have the bowels removed
excise : be removed
exhume : be removed from a grave
expunge : be removed
fillet : have the bones removed
filter (out) : be removed
free (+from;+of) : be free of
gut : have the guts removed
iron (out) : be removed
mop (+prep;particle) : be removed from
mop (up) : be removed
peel (+prep;particle) : be removed from
peel : have their peel removed
pluck : have the feathers removed
put (away) : be removed
remit : be removed
remove : be removed
rid (+of) : be free of
rinse (out) : be removed
scale : have the scales removed
scrape (+prep;particle) : be removed from

scratch (+from;particle) : be removed from
shear : have the wool removed
shell : have the shells removed
shuck : have the outer covering removed
siphon : be removed
skim : be removed
skin : have the skin removed
slip (off) : be removed
soak (particle) : be removed
sop (up) : be removed
sponge : be removed
strike (off) : be removed
strike (out) : be removed
strip (+of) : be removed of
strip : have the dress removed
take (off) : be removed
take : be removed
tease (out) : be removed
throw (off) : be removed
unburden : be relieved
uncover : have the cover removed
undress : have the dress removed
unpick : have the stitches removed
unsaddle : have the saddle removed
unscrew : have the screws removed
unseat : be removed from a position
weed : be free of weeds
wipe (+prep;particle) : be removed from

A11. Verbs that mean *to cause (something) to make a sound*

chime : chime
clack : make quick sharp sounds
clang : make a loud ringing sound
clank : make a short loud sound
clash : make a loud sound
clatter : sound a clatter
clink : make a slight high sound
crackle : make small sharp sounds
honk : make a honk
hoot : make a hoot
jangle : make a sharp sound
jingle : sound with a jingle
peal : sound loudly
pop : make a sharp explosive sound
rattle : make quick sharp sounds
ring : ring
rustle : rustle
splat : make a splat
ting : make a high ringing sound
tinkle : make light metallic sounds
toll : ring slowly and repeatedly
toot : make a warning sound with a horn
twang : sound a twang

A12a. Verbs that mean *to cause (something) to have a physical feature*

aircondition : be equipped with an air-conditioner
bevel : have a bevel
blister : form blisters
blot : have blots
breach : have an opening
bridge : have a bridge across
bruise : have a bruise
butter : have butter on it
caption : have a caption
channel : have a channel
chip : be chipped
colonize : have a colony
crease : crease
crick : have a crick
crinkle : crinkle
curtain : have a curtain
cut : have a cut
dent : have a dent
dibble : have holes
dot : have a dot
edge (+with) : have an edge of
equip : have proper equipment
feather : have feathers
flaw : have a flaw
floor : have a floor
flute : have long thin inward curves
fray : have loose threads
fuel : have fuel
furnish : have furniture
furrow : have furrows
gash : have a large deep wound
glaze : have a shiny surface
graze : have a graze
hallmark : have a hallmark
headline : have a headline
heap (+with) : have a large amount of
hedge : be surrounded by a hedge
heel : have a heel
hem : have a hem

hole : have a hole
hollow (out) : have a hollow place
illustrate : have illustrations
indent : have a dent
ink : have ink on it
label (+n;+adj) : have label
label : have a label
ladder : develop a ladder
lard : have small pieces of bacon
line : have lines
manure : have manure
mark : have marks
militarize : have military forces
nick : have a nick
notch : have a notch
number : have numbers
pattern : have a pattern
perforate : have holes
pile (+with) : have a large amount of
pit : have pits on the surface
pivot : have a pivot
pleat : have pleats
powder : have powder on the surface
puncture : get a puncture
re-cover : have a new cover
rebind : have a new binding
reface : have a new surface
reline : have a new lining
remould : have a new rubber covering
resurface : have a new surface
retread : have a new rubber covering
rewire : have new electric wires
ridge : have ridges
rifle : have grooves
rig : have necessary ropes and sails
ring : have a ring around the leg
ripple : form ripples
roof (in) : have a roof
roof : have a roof
root : form roots
rouge : have rouge
rut : have ruts
sand : have sand
sandbag : have sandbags
scallop : have scallops
scar : have a scar
seal : have a seal
snick : have a small cut
square : have straight lines and right angles
stockade : have a stockade around it
string : have strings
tag : have a tag
top (up) : have more liquid
tunnel : have a tunnel
ulcerate : have an ulcer
upholster : have comfortable coverings
weight : have weights
wound : have a wound

A12b. Verbs that mean *to cause (something) to contain something*

fluoridate : contain fluoride
garnish : contain garnish
impregnate (+with) : be impregnated with
include (+in) : be included
inject (+with) : have in the body
leaven : contain leaven
lime : contain lime
load : contain bullets
nest : be nested
poison : contain poison
salt : contain salt
spice : contain spice
spike : contain alcohol
stir (+prep;particle) : be mixed in
sugar : contain sugar
sweeten : contain some sugar or sweetness
transistorize : contain transistors

A12c. Verbs that mean *to cause (something) to be covered with something*, or *to cause (something) to cover something*

anodize : be coated with a protective film
bandage : be bound round with bandage
bedaub : be dirty with something wet and sticky
besmear (+with) : be covered with
board : be covered with boards
bury : be buried
cake (+with) : be covered with
cap (+with) : be covered with
carpet : be covered with a carpet
cloud : become covered with clouds
clutter : be cluttered
coat : be covered with
concrete : be covered with concrete

cover : be covered
creosote : be painted with creosote
deluge : be covered with water
dip : be immersed
drape (+in;+with) : be covered with
drape (+over;+round) : cover
dredge (+with;+in) : be covered with
drown : be covered with water
duck : be immersed in water
electrogalvanize : be plated with zinc
electroplate : have a coating
emulsion : be painted with emulsion paint
enamel : be covered with enamel
encapsulate : be encased in a capsule
encase : be covered completely
enshroud (+in) : be covered with
envelop (+in) : be covered completely with
fleck : be covered with flecks
flood : flood
flour : be covered with flour
fog : become covered with fog
fold (+prep;particle) : be wrapped in
frost : appear as if covered with frost
frost : become covered with frost
fur : become covered with fur
galvanize : be covered with a metal
gift-wrap : be decoratively wrapped
gild : be covered with a coat of gold
glaze : be covered with a glaze
grass : be covered with grass
gravel : be covered with gravel
grease : be covered with grease
grit : be covered with grit
heap (+on) : be piled up on
heap (+prep;particle) : be piled up
ice : be covered with icing
immerse : be immersed in a liquid
intersperse (+among;+throughout;+in) : be interspersed in
intersperse (+with) : be interspersed with
japan : be covered with a black shiny surface
lacquer : be covered with lacquer
lag : be insulated
laminate : be covered with thin metal or plastic sheets
lather : be covered with lather
line : be lined
litter (+with) : be covered with
metal : be covered with small stones
metallize : be coated or impregnated with a metal
mire : be dirty with mud
mist : become covered with mist
muck : be covered with muck
muddy : become dirty with mud
mulch : be covered with mulch
net : be covered with a net
nickel : be covered with a thin layer of nickel
oil : be coated with oil
paint : be painted
paper : be covered with wallpaper
patch : be covered or mended with a patch
pile (+onto) : cover
placard : be covered with placards
plaster (+on;+over) : stick on
plaster (+with) : be covered with
plaster : be covered with plaster
plate : be covered thinly with a metal
prepack : be wrapped up
prime : be covered with a first layer of paint
rubberize : be coated or impregnated with rubber
rust : become covered with rust
scatter (+on;+over) : be scattered over
scatter (+with) : be covered with
scatter : scatter
shower (+on) : be poured on
shower (+with) : be covered
silver : be covered with silver
slate : be covered with slates
smear (+on;+with) : be smeared on/with
smother (+with;+in) : be covered with
smudge : become dirty with a smudge
soak : be covered by a liquid
soap : be covered with soap
souse : be immersed in salted water
spatter (+prep) : be scattered on
spatter (+with) : be covered
splash (+on) : cover
splash (+with) : be covered with
sprinkle (+prep) : be scattered over
sprinkle (+with) : be covered with
sprinkle : be scattered over
steep (+in) : be immersed in
strap : have bandages tied round it
streak : be covered with streaks
strew (+over;+on) : be scattered over
strew (+with) : be covered with
stud : be covered with studs
submerge : be immersed in water
surface : be covered with a hard material
swaddle : be wrapped in many coverings
tar : be covered with tar
tarmac : be covered with tarmac
thatch : be covered with thatch
tile : be covered with tiles

tip (+with) : be covered at one end with
turf : be covered with turf
twist (+round) : be wound round
varnish : be covered with varnish
veil : be covered with a veil
veneer : be covered with a veneer
wallpaper : be covered with wallpaper
wax : be covered with wax
whitewash : be covered with whitewash
wind (+round) : be wound round
wrap (+around;+round;particle) : wrap round
wrap : be wrapped

A12d. Verbs that mean *to cause (something) to be filled with something*, or *to cause (something) to fill something*

brick (up) : be filled or enclosed with bricks
overcrowd : be filled with too many people
pack (+with) : be filled with
refill : be full again
refuel : be filled again with fuel
replenish : be filled again
saturate (+with) : be filled with
stuff (+with) : be filled with
supercharge (+with) : be filled with

A12e. Verbs that mean *to cause (something) to be decorated with (something)*

adorn : be decorated
bedeck : be decorated
blazon (+on) : be emblazoned on
blazon (+with) : be decorated
deck (+in) : be decorated
deck (+with) : be decorated
decorate (+with) : be decorated
emblazon (+on) : be emblazoned on
emblazon (+with) : be decorated
emboss (+on) : be embossed on
emboss (+with) : be decorated
festoon (+with) : be decorated
imprint (+on) : be a mark on
ornament (+with) : be decorated
panel : be decorated with panels
redecorate : have new decoration
spangle : have a shining effect; be decorated with shining objects
tattoo : be marked with a tattoo
tattoo : be tattooed on the skin
trim : be decorated

A12f. Verbs that mean *to cause (something) to have certain color(s)*

black : be black
blacken : become black
blanch : become colorless
bleach : become white
bronze : have the color of bronze
brown : become browner
char : become black
color (+adj) : have color
color : have color
colorize : be in color
crimson : become crimson
discolor : change color
dye : have a different color
dye (+adj) : have color
paint (+adj) : have color
redden : become red
stain : change in color

stain : have a stain
tan : become brown
tint : have a slight color
whiten : become white
yellow : become yellow

A13. Verbs that mean *to cause (someone) to possess (something)*, or *to cause (something) to be in the possession of (someone)*

accouter : have equipment or clothes
advance (+n) : have
advance (+to) : be in the possession of
allocate (+n) : have
allocate (+to) : be in the possession of
allot (+n) : have
allot (+to) : belong to
arm : have weapons
award (+n) : have
award (+to) : be given to
bear (+n) : have
bequeath (+n) : have
bequeath (+to) : be in the possession of
bestow (+on;+upon) : be in the possession of
build (+n) : have
bung (+n) : have
cede (+to) : be in the possession of
chuck (+n) : have
confer (+on;+upon) : belong to
deal : be in the possession of
delegate (+to) : be in the possession of
deliver (+to) : be in the possession of
empower : have the power or legal right
endow : have money
energize : have energy
enfranchise : have the right to vote
fit (out) : be equipped
fix (+n) : have
franchise : have a franchise
furnish (+with) : have
give (+to;+n) : be given to
give (back;+back_to) : be in the possession of
grant (+n) : have
grant (+to) : be in the possession of
hand (+n) : have
hand (+over_to) : have
hand (+prep;particle) : be in the possession of
invest (+with) : have
leave (+n) : have
lend (+n) : have
lend (+to) : be in the possession of
loan (+n) : have
loan (+to) : be in the possession of
mandate : have a mandate
motorize : have motor vehicles
outfit : have an outfit
permit : have
post (+n) : have
provision : have food and supplies
rearm : have weapons
rehouse : have a better home
reward : have a reward
saddle (+with) : have
sell (+n) : have
sell (+to) : be in the possession of
set (+n) : have
set (+with) : have
sling (+n) : have
supply (+to) : be in the possession of
supply (+with) : have
take (+n) : have
ticket : have a ticket
toss : have
vest (+with) : legally have
visa : have a visa

A14a. Verbs that mean *to cause (someone) to have a certain feeling or to be in a certain state of mind*

addle : become confused
affect : have feelings
aggravate : feel angry
agitate : feel anxious and nervous
ail : feel pain
alarm : be alarmed
alert : become alert
alienate : become unfriendly or unsympathetic
amaze : be amazed
amuse : be amused
anger : feel angry
annoy : be annoyed
antagonize : become an enemy

appall : be shocked
astound : be astounded
awe : feel awe
baffle : be baffled
befuddle : be confused
bias : be prejudiced
blind : be unable to understand
chagrin : feel chagrin
chill : feel fear
chloroform : become unconscious
comfort : be comforted
concuss : have a concussion
confuse : be confused
contaminate : become impure
content : feel contented
convince : be completely certain about something
cow : be cowed
crease : laugh
daunt : lose courage
daze : be dazed
delight : be delighted
demoralize : lose confidence
depress : feel depressed
desolate : be desolate
discomfit : feel uncomfortable
discompose : become worried
disconcert : be disconcerted
discontent : be discontented
disgust : feel disgust
dishearten : lose courage or hope
disillusion : free from an illusion
dismay : feel dismay
disorientate : be disorientated
displease : feel displeasure
dispose (+to) : have a feeling of; tend towards
disquiet : feel disquiet
distract : be distracted
distress : feel distress
disturb : be anxiously dissatisfied
divert : be diverted
drag (down) : feel ill
dumbfound : be dumbfounded
electrify : be greatly excited
embarrass : feel embarrassed
embitter : have bitter feelings
embolden : have more courage
enchant : be enchanted
endear (+to) : be liked by
enrage : feel very angry
enrapture : feel great joy
entertain : be entertained
enthuse : be enthusiastic
estrange : be estranged
evoke : be remembered
exasperate : feel angry
excite
excite : be excited
exhaust : tire out
exhilarate : feel cheerful
familiarize : become well informed
fatigue : become tired
faze : be surprised
fidget : fidget
flurry : be nervous and uncertain
fluster : be flustered
fray : become worn out
fret : fret
frighten : be frightened
frustrate : feel frustrated
fuddle : be in a fuddle
gall : feel angry
gladden : feel glad
grieve : grieve
gripe : feel sharp pain
habituate (+to) : become habituated to
harden : become unkind or lacking in human feelings
hearten : feel more hopeful
horrify : feel horror
humanize : be human or humane
humble : lose pride
humiliate : feel ashamed
hush : hush
hypnotize : be in a state of hypnosis
idle : idle
incense : feel very angry
infuriate : feel very angry
infuse (+into) : be filled with a quality
infuse (+with) : feel
inhibit : be inhibited
inspire (+in) : be felt by
inspire (+to) : have the desire and ability to take effective action; feel eager and confident
inspire (+with) : feel
interest (+in) : be interested in
intimidate : feel fear
intoxicate : be excited
intoxicate : become intoxicated
inure (+to) : be used to
invigorate : feel fresh and healthy
irritate : be impatient
knock (out) : be unconscious
lash : have violent feelings
lighten : become more cheerful
lull : sleep or become less active

madden : feel angry
mesmerize : be in a state of hypnosis
mystify : be mystified
narcotize : be unconscious
nark : feel angry
nettle : feel angry
nonplus : be nonplused
numb : be numb
offend : be displeased
ossify : ossify
outrage : feel outrage
overawe : be quiet because of respect and fear
overpower : become helpless
overwhelm : become completely helpless
pacify : be in peace
pacify : become calm, quiet and satisfied
pain : feel pain in the mind
panic : feel panic
peeve : feel angry and offended
perplex : be perplexed
perturb : worry
petrify : be in a state of shock
pique : feel angry
placate : be placated
play (off) : be in opposition
please : be pleased
politicize : develop an interest in or understanding of politics
prejudice : have a prejudice
provoke : feel angry
psych (out) : be frightened
psych (up) : become keen and ready
puzzle : be puzzled
quieten : become quiet
rack : feel great pain or anxiety
ravish : feel delight
reassure : be reassured
reconcile (+to) : accept
reconcile : be reconciled
refresh : be refreshed
rejuvenate : feel or look young and strong again
relax : relax
repel : have strong feelings of dislike
revolt : feel sick and shocked
rile : feel very angry
rock : be very surprised
ruffle : be upset
sadden : feel sad
scare : be scared
sedate : become sleepy or calm
shatter : be shocked
shock : be shocked
sicken : have sick feelings of dislike
silence : become silent
sober : become serious or thoughtful
solace : be comforted
spook : be afraid
stagger : feel shocked disbelief
startle : be startled
stultify : become stupid
stun : be unconscious
stun : feel surprise
stupefy : be unable to think or feel
stupefy : feel surprise
subjugate : be obedient
surprise : feel surprise
tantalize : be tantalized
tense : become tense
terrify : feel fear
thrill : feel a thrill
tire (out) : become completely tired
tire : become tired
titillate : feel excitement
torment : suffer great pain
tranquilize : become tranquil
tranquillize : become tranquil
transfix : be unable to move or think
traumatize : be deeply shocked
trouble : be worried
unbalance : become slightly mad
unhinge : become mad
unnerve : lose confidence
unsettle : feel unsettled
uplift : feel more cheerful or spiritual
vex : feel angry
weary : become weary
worry : feel anxious
worry : worry
wow : feel admiration

A14b. Verbs that mean *to cause (someone) to change his/her mind or to have certain beliefs*

brainwash : change beliefs
disabuse : not have a wrong belief
enlighten : be enlightened
inculcate (+in;+into) : be fixed in the mind of
inculcate (+with) : have
instil : be instilled in the mind of
lead (on) : believe something that is untrue
persuade : believe
sway : change their mind
talk (round) : change their mind
turn (+against) : become opposed to

A14c. Verbs that mean *to cause the body to be in a certain physical state or to experience something*

anesthetize : be unable to feel pain
asphyxiate : be unable to breathe
blind : blind
bring (+out_in) : suffer a skin condition
choke : have difficulty breathing
dazzle : be unable to see
deafen : become deaf
debilitate : become weak
enervate : become weak
enfeeble : become weak
feed (up) : become fatter and healthier
hurt : feel pain
impregnate : become pregnant
induce : give birth
inebriate : become drunk
irritate : become painful
lame : become lame
nauseate : feel nausea
overcome : become helpless
pain : feel pain
paralyse : be paralysed
prick : feel light sharp pain
prickle : feel a pricking sensation
prostrate : lose strength
smart : smart
starve : starve
tone : become stronger and healthier
torture : suffer great pain
wind : be breathless

A15. Verbs that mean *to cause (something) to increase in amount, speed, etc.*

accelerate : accelerate
accumulate : accumulate
amplify : increase
augment : increase
boost : increase
broaden : become broader
bump (up) : increase
cube : increase by cubing the amount
deepen : be deeper or increase
dilate : dilate
distend : swell
double : increase to twice the amount
elongate : be longer
embellish : be more beautiful
enhance : increase
enlarge : enlarge
enliven : be more lively
ennoble : be more noble
escalate : increase
expand : increase
extend : be longer
fan : increase
fatten : become fatter
fortify : be stronger or more effective
heighten : become higher
hike : increase
hone : become sharper
increase : increase
inflame : become more violent
inflate : inflate
intensify : become more intense
lengthen : become longer
magnify : appear larger
mark (up) : increase
maximize : increase to maximum
multiply : multiply
prolong : become longer in time
push (up) : increase
quadruple : increase four times
quicken : become quicker
raise : increase
ream : become larger
redouble : increase greatly
reflate : increase in money supply
reinforce : become stronger
rev : increase in speed
revalue : increase in value
scale (up) : increase
send (up) : increase
soup (up) : increase in power

steepen : be steeper
step (up) : increase
stimulate : become more active

strengthen : become stronger
stretch : stretch
supercharge : increase in power
swell : increase
treble : increase three times
triple : increase three times
up : increase
vulcanize : strengthen
whet : become sharper
widen : become wider

A16. Verbs that mean *to cause (something) to decrease in amount, speed, etc.*

allay : decrease
alleviate : decrease
assuage : decrease
atrophy : weaken
attenuate : decrease
beggar : become poor
blunt : decrease
blur : become less clear
brake : slow down or stop
cheapen : become cheaper
constrict : be narrower or tighter
contract : become smaller
curtail : decrease
cut (back) : decrease
damp (down) : burn more slowly
damp : sound softer
dampen : decrease
de-escalate : decrease
deaden : decrease
debase : have lower value
decelerate : decelerate
decompress : be less compressed
decrease : decrease
deflate : become smaller
defuse : be less harmful
demotivate : be less motivated
deplete : decrease
depopulate : be less populated
depress : decrease
depressurize : have less pressure
desensitize : be less sensitive
destabilize : be less stable
devalue : have less value
devitalize : have less power or strength
dilute : be dilute
diminish : decrease
downsize : be smaller
dull : become dull
ease : become less anxious
ease : become less severe
emasculate : weaken

extenuate : decrease
foreshorten : appear shorter
halve : decrease by half
knock (off) : decrease
lessen : decrease
lighten : become lighter
loosen : become less controlled
lower : decrease
minimize : decrease
mitigate : decrease
moderate : decrease
mollify : become less angry
muffle : be softer
narrow : be within a smaller range
narrow : become narrower
palliate : have less unpleasant effect
reduce (+to) : be reduced to
reduce : decrease
rein (in) : go more slowly or be less
relieve : decrease
retard : be slower
salve : be less painful
sap : decrease
scale (down) : decrease
shorten : become shorter
shrink : become smaller
shrivel : shrivel
slacken : become slack
slash : decrease
slim : decrease
slow : become slower
soft-soap : be less angry
soften (up) : become weaker
soothe : be less angry
taper : taper
telescope : become shorter
thin : become thinner
tone (down) : decrease in violence or forcefulness
truncate : be shorter

turn (down) : decrease
undermine : weaken
weaken : become weaker
whittle : decrease
wither : wither

A17. Verbs that mean *to cause (something) to improve or to be in a better state*

advance : improve
age : improve in taste
ameliorate : improve
amend : improve
beef (up) : improve
better : improve
canonize : become a saint
civilize : become civilized
civilize : become more developed
clean : become clean
cleanse : become clean
cleanse : become pure
consolidate : consolidate
cure : be healthy
cure : heal
develop : develop
elevate : be finer or more educated
elevate : have a higher rank
enrich : improve in quality
enthrone : be on the throne
fine : become pure and clear
fine : improve
foster : develop
further : advance
heal : heal
improve : improve
optimize : be as effective as possible
ordain : become a priest
polish (up) : improve
promote : have a higher rank
purify : become pure
refine : become pure
reform : improve
render (down) : become pure
touch (up) : improve
upgrade : have a higher rank

A18. Verbs that mean *to cause (something) to worsen or to be in a worse state*

abrade : wear away
adulterate : become impure
aggravate : become worse
alloy : be spoiled
bastardize : be spoiled
befoul : become foul with filth
blemish : be spoiled
blight : be spoiled
bollocks (up) : be spoiled
bugger up : be spoiled
cock (up) : be spoiled
complicate : be worse
compound : be worse
corrode : become worn away
corrupt : become morally bad
corrupt : change for the worse
cripple : be damaged or weakened
decay : decay
decompose : decompose
deface : be spoiled in appearance
defile : be impure
deform : become deformed
demote : have lower rank
deprave : be evil in character
dethrone : be dethroned
dirty : become dirty
discommode : have difficulty
disfigure : be disfigured
dislocate : be in disorder
disrupt : be in disorder
downgrade : have a lower position
erode (away) : be worn away
exacerbate : be worse
foul (up) : be spoiled
fuck (up) : be spoiled
impair : become worse
impede : slow down
impoverish : be worse
infect : become infected
injure : be injured
louse (up) : be spoiled
mess (up) : be spoiled
mire : be caught up in difficulties
muck (up) : become dirty
murder : be spoiled
mutilate : be spoiled
perish : decay
pervert : be perverted
pollute : be polluted
prejudice (+against;+in_favor_of) : become

worse
putrefy : become putrid
rot : rot
shit : become dirty
soil : become dirty
spoil : be spoiled
sully : be sullied
taint : be tainted
tarnish : tarnish
upset : be in disorder
worsen : become worse

A19. Verbs that mean *to cause (something) to be restricted in some way*

box (in) : be confined to a small space
chain (+prep;particle) : be restrained with a chain
circumscribe : be limited
confine (+to) : be limited
confine : be confined
cramp : be limited
curb : be restrained
fetter : be restricted in movement
hamper : be limited in movement
hamstring : be limited in effectiveness
handcuff : be restrained with handcuffs
hobble : be restricted in movement
immure : be confined in prison
imprison : be confined in prison
incarcerate : be confined in prison
inhibit : be restricted
intern : be confined in prison
jail : be confined in jail
keep (in) : stay inside
limit : be limited to a certain amount
localize : be limited to a small area
lock (in) : be confined to an enclosed place
pen (+in) : be confined
pen : be confined to a pen
pen : be confined to a small space
pin : be unable to move
pinion : be unable to move
quarantine : be confined in quarantine
repress : be held back
restrict : be limited to a certain amount
seal (in) : be confined to a place
tie (down) : be limited in freedom
truss : be restrained by tying up

A20. Verbs that mean *to cause (someone) to be injured*

beat (up) : be severely injured
cripple : be crippled
harm : come to harm
hurt : be injured
injure : be injured
maim : very severely wounded
run (down) : be injured
sting : be wounded or hurt
traumatize : be wounded
wing : be wounded in the arm or wing

A21. Verbs that mean *to cause (something) to become closed or blocked*

bar : be firmly closed with a bar
barricade : be closed off with a barricade
block : be blocked
blockade : be blocked
bung (up) : be blocked
clog : become blocked
close : close
cork : be closed with a cork
obstruct : be blocked up
plug : be blocked
seal (off) : close tightly
sew (up) : close
slam : shut loudly

stopper : be closed with a stopper
stuff (up) : be completely blocked
wall (up) : close with a wall

A22a. Verbs that mean *to cause (someone or something) to move*

airlift (+to) : go to
airlift : move
airmail : go by airmail
air-mail: go by airmail
attract : come near
back : go backwards
bang (prep;particle) : move violently
bear (prep;particle) : move
bounce : bounce
bounce : move with a springing movement
bring (back) : return
bring (in) : come in
budge : move a little
bundle (prep;particle) : move
bung : move
burst (+adj;+prep;particle) : move
canter : canter
capsize : capsize
cart (+prep;particle) : go
cart : move in a cart
catapult : move through the air
channel (+into) : move
chase (+prep;particle) : leave
chuck (+prep;particle) : move
chuck : move through the air
circulate : circulate
consign : move
convulse : shake violently
crank : move
dance (+prep;particle) : dance
deflect : deflect
deliver (+to) : go
derail : run off the railway line
dispatch : go
divert : change direction
download : move to another computer
drag : move
drain : drain
draw : move
drive : travel
ease (+prep;particle) : move slowly
edge (+prep;particle) : move slowly
empty (+prep;particle) : move
express : go by express mail
exude : exude
ferry (+prep;particle) : travel on a ferry
flap : flap
flick : move with a light quick sudden movement
flicker : flicker
fling : move through the air
flip : spin
float (+prep;particle) : move on the water or in the air
fly : fly
forward (+n) : go to
forward : go
freight : go as freight
frighten (+prep;particle) : go
funnel (+prep;particle) : move
gallop : gallop
get (+prep;particle) : move
goad : move
hasten : hasten
haul : move
heave (+prep;particle) : move through the air
hump (+prep;particle) : move
hurl : move through the air
hurry (+prep;particle) : go quickly
hurry : hurry
hustle : hustle
inch : move slowly
introduce (+into) : go into
jar : shake
jettison : move out
jig : move up and down
jiggle : move from side to side
joggle : joggle
jolt : jolt
launch : move
launch : move into the sky
lay (off) : leave an employment
lever (+prep;particle) : move
lift (+prep;particle) : move
lob : move through the air
loose : fly
lug (+prep;particle) : move
lure (+prep;particle) : come
maneuver (+prep;particle) : move
move : move
nose (+prep;particle) : move
nudge (+prep) : move
overturn : overturn
pack (+off_to;off) : go
paddle : move
pan : move from side to side

pipe : move through pipes
pivot : pivot
plunge (+into) : move suddenly or violently
plunge (+prep;particle) : move towards
pop (+prep;particle) : move quickly and lightly
post : go by post
precipitate (+into) : move forward forcefully
project : move through the air
propel : move forward
race (+prep;particle) : move quickly
raft : travel on a raft
recall : return
relay : go by relay
remit (+to) : go
remit : go by post
retract : retract
reverse : reverse
revolve : revolve
ripple : ripple
rock : rock
roll (back) : retreat
roll : move over and over or from side to side
rotate : rotate
route (+prep;particle) : travel
row (+prep;particle) : travel
run : move quickly or freely
rush (+prep;particle) : move suddenly with great speed
sail : travel on water
scroll : move on a screen
send (+v-ing)
send (away) : go to another place
send (in) : go to
send (off) : go
send (on) : go
send (out) : go
send : go to a place
shake : move up and down or from side to side
shift : change in position
ship (+prep) : go to
ship (+prep;particle) : go to
ship : go by ship
shoo : go away
shunt (+prep;particle) : move
shunt : move to another track
shuttle (+prep;particle) : move in shuttle
slew : slew
slide : move smoothly over a surface
sling (+prep;particle) : move through the air
snake (+prep;particle) : move in a twisting way
spin : spin
spoon (+prep;particle) : move
stampede : stampede
suck (+prep;particle) : move
suck : move into the mouth
swallow : move down the throat to the stomach
sway : sway
swing (+prep;particle) : move in a smooth curve
swing : swing
swirl (+prep;particle) : move quickly with twisting turns
swish : move quickly through the air
swivel : swivel
take (+prep;particle) : move
take : move
throw (+prep;particle) : move forcefully
throw : move rapidly through the air
tip (+prep;particle) : move
toss : move through the air
toss : toss
tow : move
transfer : move
transfer : move to another vehicle
transmit (+to) : pass to
transplant : move
transport : move
trot : trot
truck (+prep;particle) : move by truck
trundle (+prep;particle) : move
turn : change position or direction
turn : turn
twirl : twirl
twitch : twitch
unload (+from) : move
unpack : move; have the clothes removed
unsaddle : move from the saddle
uproot : leave
vibrate : vibrate
waft (+prep;particle) : move lightly on the wind or waves
waggle : waggle
walk (+prep;particle) : move like walking
wheel : move
whip (+prep;particle) : move quickly
whirl (+prep;particle) : move round and round very fast
whirl : move in a hurry
whisk (+prep;particle) : move quickly
whisk : move quickly

wiggle : move from side to side
withdraw : withdraw
wobble : wobble
worm (+into;particle) : move
zap (+prep;particle) : move quickly
zip (+prep;particle) : move quickly

A22b. Verbs that mean *to cause (someone or something) to fall or move down*

airdrop : drop from an aircraft
bowl (over) : fall down
bring (down) : come down
cast : drop
chop (down) : fall down
cut (down) : fall down
dip : dip
down : fall down
drip : drip
drop : drop
fell : fall down
floor : fall down
get (down) : come down
lower : go down
overbalance : overbalance
parachute (+prep;particle) : drop from an aircraft
parachute : parachute
poleaxe : fall down
prostrate : be in a prostrate position
rain (+prep;particle) : fall like rain
send (down) : go down
take (down) : come down
tip : tip
topple : topple
trip : trip
unhorse : fall from a horse
unseat : fall from the saddle
upend : fall down
upset : fall over

A22c. Verbs that mean *to cause (something) to come out*

dart (+prep;particle) : go out suddenly and quickly
dig (out) : come out
draw (out) : come out
fish (out) : come out
gouge (out) : come out
jet : jet out
out : go out
pluck : come out
pour (+prep;particle) : flow out of or into
shed : flow out
slop : spill
spew (+prep;particle) : come out in a rush
spill : spill
spout : come out in a forceful stream
spray : come out in small drops
spurt : spurt
squirt : come out in a thin fast stream

A22d. Verbs that mean *to cause (something) to rise or move up*

boost : rise
fish (up) : come up
hitch (up) : come upwards
hoist : rise
jack (up) : rise
levitate : rise and float in the air
lift : rise to a higher level
raise : rise
run (up) : rise
stand : be erect
uplift : rise high
winch : rise

A22e. Verbs that mean *to cause (someone or something) to be located at a certain place*

bag : be placed in a bag
bank : be kept at the bank
beach : be placed on the beach
berth : come into a berth
bottle : be placed in bottles
bottle : be preserved in bottles
box : be placed in a box
bundle (prep;particle) : be stored

cage : be kept in a cage
can : be preserved in cans
center : be located in the center
clap (prep;particle) : be placed in
consign (+prep) : be located
containerize : be packed in containers
corral : be located in a corral
cram (+prep;particle) : be crammed in
crate : be packed in a crate
deposit (+prep;particle) : be located
deposit (+prep;particle) : fall and remain
dispose (+prep;particle) : be placed on
dock : be located at a dock
drop (prep;particle) : be located
enclose : be inside an envelope
entomb : be placed in a tomb
entrain : be placed on a train
file : be placed in a file
garage : be kept in a garage
inject (+into) : be in the body
input (+into) : be stored
insert : be inserted
inset : be placed as an inset
install (+prep;particle) : be located
institutionalize : be in an institution
interpose (+between) : be placed between
jam (+prep;particle) : placed tightly in
kennel : be kept in a kennel
land : land
lay (+prep;particle) : be located
locate (+prep;particle) : be located
lock (away) : be kept in a secure place
lock : be kept in a safe place
lodge (+prep;particle) : be located
maroon : be marooned
nestle (+prep;particle) : settle
pack (+prep;particle) : be located
package : be placed in a package
park (+prep) : be located
park (+prep;particle) : be placed
pasture : be left in a pasture to feed
perch (+prep;particle) : perch on
place (+prep;particle) : be located
plant (+prep;particle) : be located
plant (+prep;particle) : be placed firmly on
plonk (+prep;particle) : be placed on
pocket : be placed in the pocket
poise (+prep;particle) : be placed on
position (+prep;particle) : be located
post (+prep) : be located
post (+prep;particle) : be posted
post : be posted
pot : be located in a pot filled with earth
put (+prep;particle) : be located
recess : be located in a recess
relegate : be located in a worse place
relocate : be located in a new place
repose (+on;particle) : be located
sandwich (+prep;particle) : be placed between
scratch (+on;particle) : be put on
set (+in) : be located
set (+on) : attack or chase
set (+prep;particle) : be located
settle (+prep;particle) : be located
settle : settle
sheathe : be kept in a sheath
shelve (+prep) : be located
shove (+prep;particle) : be located
sit (+prep;particle) : be located
site (+prep;particle) : be located
situate (+prep;particle) : be located
slam (+prep;particle) : be located
sling (+prep;particle) : to hang
slot (+prep;particle) : be placed in a slot
squash (+prep;particle) : be located
stable : be kept in a stable
stand (+prep;particle) : be located
station (+prep;particle) : be located
store (+prep) : be kept
stow (+prep) : be kept
string : be placed on a thread
tin : be packed in tins
tuck (+prep;particle) : be located

A22f. Verbs that mean *to cause (something) to hang from some place*

dangle : dangle
drape : hang loosely
hang : hang
loll : hang loosely

string (up) : hang high
suspend (+from;particle) : hang from

A23. Verbs that mean *to cause (someone or something) to be become the thing specified*

anoint (+n) : become
baptize (+n) : become
canalize : be like a canal
create (+n) : become
criminalize : become a criminal
crown (+n) : become
curry : become curry
dub (+n) : become
ennoble : become a nobleman
enslave : become a slave
fanaticize : become a fanatic
federate : become a federation
fossilize : become a fossil
install (+as) : be
institutionalize : become an institution
knight : become a knight
loop : become a loop
magnetize : become a magnet
make (+n) : become
malt : become malt
martyr : become a martyr
mineralize : become an ore or mineral
mummify : become a mummy
orphan : be an orphan
outlaw : be an outlaw
parcel (up) : become a parcel
queen : become a queen

A24a. Verbs that mean *to cause (something) to be joined or connected*

concatenate : join together
connect : be connected
couple : be joined together
dovetail : be joined
engage : engage
fuse : become joined
hook (up) : be connected to a power supply or central system
integrate : be joined
interlace : be joined
interlink : be joined
interlock : interlock
join : be joined
knot : be joined together
link : be connected
network : be connected
plug (in) : be connected to a power supply
put (through) : be connected by telephone
rejoin : be joined again
solder : be joined or repaired with solder
splice : be joined together end to end
weld : be joined
yoke : be joined

A24b. Verbs that mean *to cause (something) to be fastened*

bind : be fastened
bind : be tied together
bond : stick together
buckle : fasten with a buckle
button : fasten with buttons
clasp (+prep;particle) : be fastened
clip (+prep;particle) : be fastened
embed (+in) : be embedded in
fasten : be fastened
fix (+prep;particle) : be fastened on
gird : be fastened
glue : stick
gum (+prep;particle) : stick
harness : be fastened to
hinge : be fixed on hinges
hitch (+prep;particle) : be fastened to
hook (+prep;particle) : be fastened on
lace : be fastened
lash (+prep;particle) : be fastened
latch : be fastened
lock : be fastened
loop : be fastened
moor : be fastened
nail (+prep;particle) : be fastened on
paste (+prep;particle) : be stuck on
peg : be fastened with a peg
pin (+prep;particle) : be fastened with pins
remount : be fastened on a new piece of cardboard
rivet : be fastened with rivets
rope (+prep;particle) : be fastened
screw (+prep;particle) : be fastened with

screws

seal : be fastened
secure : be fastened
sellotape : be mended with sellotape
sew (+prep;particle) : be fastened by stitches
stake : be fastened to stakes
staple : be fastened with staples
stick (+prep) : stick
strap (+prep;particle) : be fastened
tack : be fastened with a tack
tape : be fastened with tape
tether : be fastened with a tether
tie (+prep;particle) : be fastened
tie : be fastened
wedge : be firmly fixed
wire : be fastened with wires

A24c. Verbs that mean *to cause (something) to be twisted together*

braid : be braided
entangle : become entangled
entwine : twist together, round or in
intertwine : intertwine
interweave : weave together
plait : be in plaits
ravel : ravel
tangle : tangle
twine : twist or wind
twist (together) : be twisted together
weave (+prep;particle) : be twisted or wound
wind (+into)

A25a. Verbs that mean *to cause (something) to be unfastened*

disconnect : be disconnected
disengage : come loose and separate
loose : be unfastened
unbar : be unfastened
unbind : be unfastened
unbuckle : be unfastened
unbutton : be unfastened
unchain : be free
unclip : be unfastened
uncouple : separate
undo : be unfastened
unfasten : be unfastened
unhook : be unfastened
unlatch : be unfastened
unlock : be unfastened
unloosen : become loose
unscrew : be undone
untie : be unfastened
unwind : be unwound
unzip : be unfastened

A25b. Verbs that mean *to cause (something) to open or to be opened*

open (up) : open
open : open
reopen : open again
throw (open) : be open to the public
unbar : be open
uncork : be open
unfold : open
unfurl : open
unplug : be disconnected; be opened by removing a plug
unroll : open
unseal : open
unstop : be open

A26a. Verbs that mean *to cause (something) to separate or break into smaller pieces*

amputate : be separated from the rest
balkanize : separate into smaller units
bisect : be separated into two parts
blast (+prep) : break up
break : break
chip : be in small pieces
chop (up) : be in small pieces
cleave : separate
cream : be reduced to a thick soft mixture
crumble : break into very small pieces
crush : break into a powder
cube : be in small cubes
curtain (off) : be separated
cut (+prep;particle) : be separated
cut (off) : be separated
cut (out) : be separated
cut (up) : be in little pieces
decollectivize : no longer be a collective
decompose : separate into parts
demobilize : leave military service
departmentalize : separate into departments
detach : be separated
dice : be in small square pieces
diffract : separate into the spectrum
dig : break up
disperse : disperse
dissipate : dissipate
dissolve : dissolve
divide : separate into groups or parts
explode : explode
extract : be taken out
fence (off) : be separated
fragment : break into fragments
grind : be reduced to small pieces or powder form
halve : be separated into halves
have (out) : be taken out
isolate : be isolated
leach : separate
liquidize : be reduced to a liquid form
lobotomize : have a lobe of the brain separated
mangle : be torn to pieces
mark (off) : become a separate area
mash : be reduced to a soft substance
mince : be in very small pieces
modularize : separate into modules
part : separate
partition : be separated into parts
partition (off) : become separate
plough : break up and turn over
polarize : be separated into groups
pound : be reduced to a soft mass or powder
pulverize : be reduced to a powder form
puree : be reduced to a puree
quarter : be separated into four parts
regionalize : separate into regions
rend : split
rip (up) : be torn into pieces
rip : be torn
rope (off) : be separated
scarify : break up
sectorize : be separated into sectors
segment : be separated into segments
separate : separate
sever : be separated
shatter : break suddenly into very small pieces
shred : be torn into shreds
slice : be in slices
smash : break into pieces violently and noisily
snap : break suddenly and sharply into two parts
splinter : break into small sharp-pointed pieces
split : separate into parts
split : split
square : be divided into squares
subdivide : be divided
sunder : sunder
take (apart) : separate into pieces
tear : be torn
tear (+prep;particle) : be torn
touch (off) : explode
unravel : unravel
uproot : be torn out of the earth
wall (off) : be separated
zone : be separated into zones

A26b. Verbs that mean *to cause (something) to be physically damaged*

break : break
burst : burst
bust : break
crack : break
crack : break
crash : crash
crash-land : crash
crumple : crumple

crush : break

damage : be damaged
fracture : fracture
lacerate : be torn
mutilate : be seriously damaged
rupture : rupture
squash : be squashed
stave (in) : break inwards
vandalize : be damaged

A27. Verbs that mean *to cause (someone) to be set free from something*

bail (out) : be released
disentangle : be free
emancipate : become free
enfranchise : be free
exorcize : leave, be free of an evil spirit
extricate : be free
free : be free
free : be set free
liberate : be free

parole : be free on parole
ransom : be set free
release : go free
rescue : be set free
unleash : be free from control
unloose : be unrestrained
unshackle : be free
untangle : be untangled

A28. Verbs that mean *to cause (something) to be safe*, or *to protect (something)*

immunize : be immune against disease
protect : be safe
safeguard : be free or safe
save : be safe
screen : be sheltered or protected
secure : be safe
shade : be sheltered from light or heat
shelter : be sheltered
vaccinate : be immune

A29. Verbs that mean *to cause (an event) to be delayed*

defer : be delayed
delay : be delayed
hold (over) : be delayed
hold (up) : be delayed
postpone : be delayed
put (back) : be delayed
put (off) : be delayed

A30. Verbs that mean *to cause (someone) to lose something*

deprive (+of) : lose
dispossess : lose their property
do (+out_of) : lose
fool (+out_of) : lose

lose (+n) : lose
relieve (+of) : be free of

A31. Verbs that mean *to cause (people) to gather, unite or form a group*

aggregate : gather into a group or mass
assemble : gather
bunch : gather into bunches
cartelize : form a cartel
clump : gather into a clump
cluster : gather or grow in clusters
collect : gather to form a group or mass
collectivize : form a collective
combine : unite
concentrate (+prep;particle) : gather in one place
confederate : combine in a confederacy
conflate : combine
conjoin : unite
consolidate : merge
convene : meet together
convoke : meet together
drift : pile up
factionalize : form factions
federalize : unite under a federal government
group : gather into groups
herd (+prep;particle) : gather in a large group
league : unite
merge : merge
muster : gather
parade : gather together in a parade
rally : rally
re-form : form again
regroup : gather into groups again
reunite : unite again
round (up) : gather
stack : form a stack
syndicate : form into a syndicate
unify : unite
unite : unite

A32. Verbs that mean *to cause (someone) to wear something*

attire : wear
bridle : wear a bridle
doll (up) : be dressed up
dress (+prep) : wear
dress : wear clothes
garland : wear a garland
harness : wear a harness
muzzle : wear a muzzle
rig (+out) : wear special clothes
robe : wear a robe
saddle : wear a saddle
shoe : wear a horseshoe
slip (on) : be worn
throw (on) : be worn

A33. Verbs that mean *to cause (something) to be put right or to be back in good working order*

correct : be set right
mend : be repaired
recondition : be back in working order
reconstitute : be back in its former condition
rectify : be set right
redeem : be back in favor
redress : be set right
rehabilitate : be back in good condition
rehabilitate : be rehabilitated
reinstate : be back in a former position
remedy : be set right
renovate : be back in good condition
repair : be set right
repair : work again
reset : be back in place again
restore : be back in a former situation
restore : be back in existence
resurrect : be back in existence
right : be upright again
straighten (out) : be set right

A34. Verbs that mean *to cause (someone or some animal) to be castrated*

alter : be castrated
castrate : be castrated
doctor : be unable to breed
emasculate : be castrated
geld : be castrated
spay : be castrated

A35. Verbs that mean *to cause (something) to be legal*

enact : become law
legalize : become legal
legitimize : be legitimate
legitimize : become legal or acceptable
ratify : be official
regularize : become legal and official
validate : become valid

A36. Verbs that mean *to cause (something) to change physically or chemically*

atomize : become minute particles or a fine spray
carbonize : change into carbon
clot : form into clots
coagulate : solidify
condense : become liquid
congeal : become thick or solid
crystallize : form crystals
curdle : form into curd
degrade : change to a lower or simpler kind
distil : be distilled
emulsify : become an emulsion
evaporate : vaporize
ferment : ferment
freeze : harden
ionize : ionize
liquefy : become liquid
melt (down) : melt
melt : melt
ossify : change into bone
petrify : turn into stone
plasticize : become plastic
pulp : become pulp
smelt : melt
solidify : become solid
sublimate : change into gas and back to solid
thaw : thaw
transmute : change into another substance
vaporize : vaporize
vitrify : change into glass

A37. Verbs that mean *to cause (something) to change in some unspecified way*

acculturate : change
adapt : adapt
adjust : change
alter : become different
amend : be changed
change : change
convert (+to;+into) : change into
fluctuate : fluctuate
metamorphose : metamorphose
modify : change slightly
remodel : change shape
revamp : have a new form or structure
revolutionize : completely change

skew : be not straight or not exact
transform : change completely
transmogrify : change complete
turn (+prep;particle) : change to

A38. Verbs that mean *to cause (something) to be aligned or arranged in a particular way*

aim (+at) : aim at
align (+with) : come into the same line as
align : become aligned
angle : be at an angle
arrange : be in a particular order
array : be set in order
cant : slope or lean
cock : be erect
disarrange : be untidy
dislocate : be dislocated
disorder : be in a state of disorder
invert : be inverted
jumble : be mixed in disorder
muddle : be in disorder
muss : untidy
permute : be in a different order
picket (+prep;particle) : be in position as pickets
plaster (+prep;particle) : lie flat or stick to another surface
point (+at;+towards) : point at
rearrange : be in a different order
rumple : become disarranged
slant : slope
string (out) : be in a line
target (+at;+on) : aim at
tilt : tilt
transpose : be in reverse order
tousle : be untidy
twist : twist

A39. Verbs that mean *to cause (something) to have a different shape*

arch : become an arch
bend : become bent
buckle : become bent
coil : wind round
crook : be bent
curl : form curls
curve : curve
distort : be distorted
double : be folded in half
fold : be folded
frizz : go into tight short curls
furl : fold up
perm : have a perm
straighten : become straight
unbend : become straight
warp : warp

A40. Verbs that mean *to cause (something) to be revealed or uncovered*

conjure : appear
disinter : be uncovered
materialize : materialize
reveal : be seen
unearth : be uncovered
unfold : unfold
unsheathe : uncovered or removed from a sheathe
unveil : be uncovered
unwrap : be revealed

A41. Verbs that mean *to cause (something) to be concealed or hidden from view*

blot (out) : be difficult to see
blur : become difficult to see
bury : be buried

conceal : be hidden
obliterate : be not seen
screen : be hidden from view
secrete : be hidden
submerge : be covered or hidden

A42. Miscellaneous causal verbs where the effect can be described with an adjective

accustom : become accustomed
acidify : become acidic
actualize : become actual
age : become old
americanize : become American
anglicize : become English
anneal : become hard
bake : become cooked
bake : become hard
bankrupt : become bankrupt
barbarize : become cruel
beautify : become beautiful
blow-dry : become dry
bone : become stiff
brighten : become bright
broil : be very hot
brutalize : become brutal
bureaucratize : be bureaucratic
calcify : become hard
calm : become calm
cement : be firm
chafe : become sore or worn
chap : become sore and cracked
chill : become cold
clarify : become clear
clean (out) : clean and tidy
clear : become clear
cloud : become less clear
coarsen : become coarse
complete : be complete
condense : become thicker
consecrate : become holy
cool : become cool
crisp : become crisp
damp : be damp
dampen : become damp
darken : become dark
decolonize : be politically independent
dehydrate : be completely dry
delegitimize : become legitimate
democratize : become democratic
demystify : less mysterious
desiccate : dry
dim : become dim
disconnect : be disconnected

drain : drain
drench : be thoroughly wet
dry : be dry
dry : become dry
elasticize : be elastic
empty : become empty
enrich : become rich
etherialize : be ethereal
falsify : become false
fill (+adj) : be
fireproof : become fireproof
firm : become firm
flatten : become flat
flatten : sound flat
freeze-dry : be dry
fuzz : be fuzzy
globalize : become global
harden : become hard
heat : become hot
historicize : become historic
homogenize : be homogeneous
humidify : become humid
ice : become cold
illegalize : be illegal
immobilize : be immobile
impoverish : become poor
irradiate : become bright
jazz (up) : become more active, interesting or enjoyable
jolly (up) : become bright and cheerful
level : become flat
liberalize : become liberal
lighten : become brighter
loosen : become loose
macerate : become soft
marginalize : be marginal
marry (off) : be married
mature : become mature
mellow : become mellow
modernize : become modern
moisten : become moist
moisturize : be moist
neaten : be neat
normalize : become normal
oxidize : become rusty

passivize : become passive
pauperize : become poor
perfect : become perfect
plump (up) : become rounded and soft
polarize : be polarized
polish : be polished
popularize : become popular
preheat : be hot
prettify : become pretty
protract : last longer in time
radicalize : be radical
rationalize : be more modern and effective
ripen : ripen
roboticize : be automatic
robotize : be automatic
roughen : become rough
round : become round
rub (+adj)
ruralize : be rural or have a rural appearance
rustproof : be rustproof
sanctify : holy
sand : be smoother
sanitize : be less unpleasant
saturate : be wet
scorch : be dry
secularize : become secular
send (+adj;+prep;particle)

sensitize : be sensitive
set (+adj;particle)
shadow : be dark
shake (+adj) : become dry
sharpen : become sharp
simplify : become simpler
sleek : become sleek
slick (down) : become smooth and shiny
smear : be unclear
smooth : become smooth
soften : become soft
soundproof : become soundproof
sour : become sour
stabilize : become stable
steady : become steady
sterilize : be sterile
stiffen : become stiff
still : become still
streamline : be more simple and effective
supercool : cool below freezing point without solidifying
tart (up) : become more attractive
tauten : become taut
tenderize : become tender
thicken : become thick
tidy : become tidy
tighten : become tight
toast : become warm
toughen : become tough
turn (+adj)
unify : be uniform
unravel : become clear
update : become modern or up-to-date
warm (over) : be warm
warm (up) : be warm
warm : become warm
waterproof : become waterproof
weatherproof : become weatherproof
wet : become wet
zip (+adj) : be

A43. Other miscellaneous causal verbs.

acclimatize : acclimatize
advantage : have an advantage or benefit
afforest : be planted with trees
alienate : change ownership
attach (+to) : a member of a group
automate : be automated
balance : balance
boil : boil
calibrate : be calibrated
christen (+n) : have the name
circulate : circulate

civilianize : be under civilian control
compost : become compost
compress : be compressed
computerize : use a computer
contort : contort
cross-fertilize : be fertilized
crossbreed : crossbreed
decentralize : be decentralized
decide : decide
declassify : be not a secret anymore
defeminize : be without feminine

characteristics
demilitarize : not to have a military character
denationalize : be not owned by the government
denuclearize : be without nuclear armaments
depoliticize : not have a political character
detonate : detonate
diffuse : diffuse
disguise : change the appearance or character
disqualify (+for;+from) : be unsuitable or unable to
disunite : be in a state of disunity
diversify : be diversified
domesticate : be domesticated
domesticate : be domesticated
dose : take medicine
dovetail : dovetail
dwarf : not grow properly
electrify : be run by electricity
embarrass : have difficulties with money
embroil : be embroiled
endanger : be in danger
enlist : enlist
entitle (+n) : have title
excommunicate : be no longer a member of the church
exercise : exercise
expedite : go faster
facilitate : be easier
fade : fade
feminize : have feminine characteristics
fertilize : be fertilized
fight (off) : keep away
flare : flare
flavor : have a flavor
float : float
floodlight : be lighted
fluidize : behave like a fluid
focus : be focused
foist (+on) : be suffered for a time by
fructify : produce fruit
fructify : produce successful results
germinate : germinate
gild : have an attractive appearance
graft (+onto) : put into a body as a graft
graze : graze
ground : ground
harmonize : harmonize
hatch : hatch
hatch : hatch
head (off) : change direction
hit (+on;+against) : hit something
import : come from another country
industrialize : industrialize
inflect : change in level
insinuate (+into) : be accepted into something
integrate : integrate
interbreed : interbreed
internationalize : be under international control
juice (up) : have more life or excitement
lend (+n) : have a quality of
lend (+to) : be a quality of
let (+in_for) : experience
let (+in_on) : share a secret
light : be lighted
mandate : be under a mandate
mate : mate
mete out (+to) : be suffered by
militarize : have a military character
mould : fit closely
name (+n) : have the name
naturalize : live in a new place
obfuscate : be difficult to understand
obtrude : obtrude
obtrude : stick out
outface : look away
overburden : carry too much or do too much work
overdevelop : be overdeveloped
package : be in the form of a package
pair : pair
parch : parch
pep (up) : be more active or interesting
percolate : percolate
perfume : have a pleasant smell
perfume : have a pleasant smell
politicize : have a political character
pose : pose
privatize : be not owned by the government
project : stick out
pucker : tighten into uneven folds
quicken : come to life
ramify : ramify
reafforest : be planted with trees
reecho : repeat again
reelect : be elected again
refract : change direction
rehearse : rehearse
reincarnate : return to life in a new body
rename : have a new name
reset : show a different number
rest : rest on
rest : take a rest
retire : retire

reverse : change position
scald : be almost boiling
scorch : be burnt
sculpture (+into)
scupper : sink
scuttle : sink
shipwreck : suffer shipwreck
short-circuit : short-circuit
sidetrack : be sidetracked
sink : sink
splash : splash
splay : spread out
spread : spread
sprout : sprout
stalemate : be in a stalemate
standardize : be standard
starch : stiffen
stare (out) : look away
starve (+of) : lack
stifle : stifle
stink (out) : fill with a bad smell
strangulate : strangulate
streamline : be streamlined
stretch : reach full width or length
subject (+to) : experience
superheat : heat beyond boiling point without vaporizing
synchronize : happen at the same time or speed
synchronize : show the same time
talk (out) : be settled
tan : change into leather
taper : taper
throw (+back_on) : have to depend on
toast : be toasted
transfigure : change appearance
trip : make a mistake
unionize : become a member of a trade union
upend : stand on end
urbanize : have urban characteristics
vote (+prep;particle)
waste : lose flesh and strength gradually
water (down) : be diluted with water
wilt : wilt
wind : tighten
work : do work
wreck : be in a shipwreck

B. Verbs that mean *to be caused by something*

proceed from
result from
stem from

C. Verbs that mean *to prevent something*

C1. Verbs that mean *to prevent (an event)*, or *to prevent (something) from coming into existence*

(Note: This is to be distinguished from *to stop something happening*. *To stop something* means to cause an ongoing event to come to an end. *To prevent something* means to cause something that would otherwise happen to not happen.)

avert
avoid
beat (off)
call (off)
cancel
choke (off)
cover (up) : from being noticed
disappoint : from being fulfilled
forestall
frustrate : from being fulfilled
gum (up) : from working properly
hide : from being seen or found
hold (back) : from moving forward
hold (down) : from rising
hush (up) : from being known
keep (down) : from increasing
preclude
preempt
prevent
prohibit
repel : from succeeding
repulse : from succeeding
rescue (+from_v-ing)
screen (out) : from coming in
smother
stave (off)
stifle
stonewall
stunt : from growing fully
stymie
thwart
veto
ward (off)

C2. Verbs that mean *to prevent or stop (someone) from doing (something)*

bar (+prep;particle) : from coming in or going out
crowd (out) : from coming in
debar (+from)
debar (+n) : from performing
deny (+n) : from having
detain : from leaving
deter : from acting
foil : from succeeding
gag : from speaking
ground : from flying or going out
hinder (+from)
hold (off) : from advancing
inhibit (+from)
lock (out) : from entering
nobble : from winning
prevent (+v-ing)
prohibit (+from)
restrain (+from)
silence : from expressing opinions or making opposing statements
stop
stop (+from)
stop (+v-ing)

C3. Verbs that mean *to persuade (someone) not to (do something)*

bamboozle (+out_of)
cajole (+out_of)
coax (+out_of)
con (+out_of)
dissuade
jolly (+out_of)
persuade (+out_of)
reason (+out_of)
shame (+out_of)
talk (+out_of)

D. Verbs that mean *to affect (something)* without specifying in what way

act on/upon
affect
condition
impact
impinge on

APPENDIX 4. LINGUISTIC PATTERNS FOR IDENTIFYING CAUSE AND EFFECT EXPRESSED IN TEXT 170

Appendix 4

170

Contents

A. Notation Used in the Patterns

A pattern consists of a sequence of tokens separated by spaces. Each token is one of the following:

(1) an "&" sign followed by a *subpattern* label, denoting a set of subpatterns. The set of subpatterns denoted by each subpattern label is listed in a *subpatterns file*. The two subpattern files used in this study are listed in Section F of this appendix. Each subpattern is also a pattern, i.e. a subpattern is a sequence of tokens as described in this section. A set of subpatterns may include an empty subpattern as one of the subpatterns. An empty subpattern is indicated in the subpatterns file by the underscore character "_".

(2) an "@" sign followed by a *part-of-speech* label. Part-of-speech labels are listed in Appendix 2, Section A. The pattern matching program assumes the part-of-speech label to be truncated. For example, the token "@V" in a pattern will match any of the following part-of-speech labels in the text:

VB	base form verb
VBD	past tense verb
VBG	gerund or present participle verb
VBN	past participle verb
VBP	non-3rd person singular present tense verb
VBZ	3rd person singular present tense verb

(3) an "$" sign followed by a *verb group* label, denoting a set of verbs. Each set of patterns is associated with a *verb groups file* that lists the verbs for each verb group. The symbol "|" followed by a part-of-speech label may be added to the end of the token, indicating that the token will match a word belonging to the specified verb group and having the specified part-of-speech label. Section D below describes the kinds of verb groups used in this study.

(4) a single asterisk "*". This is a wild card symbol indicating at most one word, i.e. the token will match zero or one word. For example, the pattern "word1 * word2" will match the phrase "word1 word3 word2" (which has one intervening word between "word1" and "word2") as well as the phrase "word1 word2" (where "word1" and "word2" are adjacent).

(5) three asterisks "***". This is a wild card symbol indicating any number (including zero) of adjacent words. For example, the pattern "word1 *** word2" will match any phrase that begins with "word1" and ends with "word2" with any number of words in between.

(6) "[1]". This token is a wild card symbol indicating one or more adjacent words. It is similar in effect to the token "***", except that it must match with at least one word, and the words that it matches cannot consist solely of punctuation marks. This token also specifies the side effect that if the pattern succeeds in matching a sentence, the words that match with this token are identified as the *first member* of the relation. (The *first member* of the causal relation is the *cause*, and the *second member* is the *effect*.) This explains why the token is not allowed to match just punctuation marks: punctuation marks by themselves do not mean anything and thus cannot represent the *cause* in a causal relation.

(7) "[2]". This token has the same effect as "[1]" except that the words that match with this token are identified as the *second member* of the relation, i.e. the *effect*.

(8) "[X:1]" where *X* is one of the phrase labels listed in Appendix 2, Section B. This token has the same effect as "[1]" except that only a phrase of the type *X* is allowed to match with this token. For example, "[N:1]" indicates that only a noun phrase can match with this token, and the noun phrase is identified as the *first member* of the relation if the pattern succeeds in matching the sentence.

(9) "[x:1]" where *x* is one of the part-of-speech labels listed in Appendix 2, Section A, converted to lowercase characters. This token has the same effect as "@X" except that the word that matches with this token is identified as the *first member* of the relation. For example, "[v:1]" will match any one of the part-of-speech labels *VB, VBD, VBG, VBN, VBP* and *VBZ* (part-of-speech labels for the various verb forms), and the matching verb is identified as the *first member* of the relation.

(10) "[X:2]" where *X* is one of the phrase labels listed in Appendix 2, Section B. This token has the same effect as "[X:1]", except that the phrase that matches with this token is identified as the *second member* of the relation.

(11) "[x:2]" where *x* is one of the part-of-speech labels listed in Appendix 2, Section A, converted to lowercase. This token has the same effect as "[x:1]", except that the word that matches with this token is identified as the *second member* of the relation.

(12) "[X:0]" where *X* is one of the phrase labels listed in Appendix 2, Section B. This token will match a phrase of the type *X*. For example, "[N:0]" will match a noun phrase. This token has the same effect as "[X:1]" and "[X:2]", except that the phrase that matches with this token is *not* identified as the first or second member of the relation, i.e. no side effect is specified by this token.

(13) any token not belonging to one of the above 12 types is a *literal*. A *literal* token will match any word or punctuation mark that has the same characters as the token. A *literal* token may be terminated by one of the following:
- the truncation sign "*". For example, the token *retriev** will match the words *retrieve*, *retrieves*, *retrieved*, *retrieval* and *retrieving*.
- the "|" sign followed by a part-of-speech label. For example, the token "fight|V" will

match the verb "fight|VB" but not the noun "fight|NN".

- the truncation sign "*" followed by the "|" sign followed by a part-of-speech label. For example, *retriev*|V* with match the words *retrieve|VB*, *retrieves|VBZ*, *retrieved|VBD*, *retrieving|VBG*, but not *retrieval|NN*.

Literal tokens must be in lower case unless they are meant to match proper nouns, in which case they must be in upper case.

I refer to tokens of types (6) to (11) as *slots*. They are like blanks in a template for one to fill in the *cause* or *effect*.

Any type of token may appear in a pattern any number of times and in any position. However, a pattern must contain at least one token which is not a punctuation mark and which belongs to either type (2), (3) or (13). The reason for this constraint is explained in Section B below.

Finally, tokens of type (2), (3) and (13) may be terminated by the symbol "~". The effect of this symbol is also explained in Section B.

B. Procedure for Pattern Matching

A pattern specifies a sequence of linguistic items for the pattern matching program to look for in a document. The patterns listed in Sections E1 and E3 are constructed to match text (i.e. words and phrases) within a sentence. The pattern matching program looks within each sentence for the sequence of items specified by these patterns. The patterns listed in Sections E2 are constructed to match text across a sentence boundary. They are meant to be applied to pairs of adjacent sentences.

It is possible for a pattern to match a sentence (or pair of adjacent sentences) in more than one way. This is because the wild card symbol "***" and the slots "[1]" and "[2]" can match an arbitrary number of words. In addition, some tokens represent a set of alternative subpatterns to try. The order in which the alternatives are tried determines how the pattern will match the sentence. This section describes the order in which alternatives are tried for each type of token and the general procedure used to match patterns with sentences.

Before pattern matching, the sentences are processed in the following ways:

- each word (and punctuation mark) is tagged with one of the part-of-speech labels listed in Appendix 2, Section A.
- phrases are bracketed and tagged with phrase labels listed in Appendix 2, Section B.

A sample of text that have been tagged with part-of-speech labels and phrase brackets are given in Appendix 2, Section C.

Before attempting to match a pattern with a sentence (or pair of adjacent sentences), the pattern matching program first adds the wild card symbol "***" to the beginning of the pattern. The effect of this is that the pattern need not match the whole sentence. It needs to match just one part of the sentence.

The pattern matching program also adds a period to the beginning of the sentence being processed. The effect is that the period now marks the beginning of a sentence. This makes it possible for a pattern to specify that something must appear at the beginning of the sentence. The period is, of course, also the punctuation mark at the end of a sentence (unless the sentence is terminated by some other punctuation mark, e.g. "?"). So, if a period appears as a token in a pattern, this indicates either the beginning or end of a sentence.

Pattern matching is assumed to proceed from the beginning of the pattern to the end. The pattern matching program "consumes" (i.e. removes) the first token from the pattern, and then checks whether the item indicated by this token occurs at the beginning of the sentence. If so, the pattern matching program "consumes" this item from the beginning of the sentence. The token that is consumed from the pattern is said to match the item consumed from the sentence.

The program then consumes the first token from the remaining pattern (i.e. the tokens that still remain in the pattern), and checks whether the item indicated by this token occurs at the beginning of the remaining text (i.e. what is left of the sentence at this point in the processing). If found, the program consumes this item from the remaining text. This step is iterated until all the tokens in the pattern have been consumed and all the linguistic items indicated by the tokens have been found and consumed from the sentence. When this happens, the pattern succeeds in matching the sentence. (Note: It is not necessary that all the words in the sentence be consumed for a pattern to succeed. It is sufficient that all the tokens in the pattern are consumed.)

Some tokens (types 1, 4, 5, 6, 7, 8, 10, and 12 described in Section A above) specify alternative items for the pattern matching program to look for. When it is possible for more than one alternative to match the beginning of the remaining text, the order in which these alternative items are tried by the program can affect the result of the pattern matching. Also, if a subsequent token fails to match the beginning of the remaining text, the program can backtrack to this token and try the other alternatives. The following describes the order in which the alternatives are tried for each type of token:

- For a type 1 token (i.e. an "&" sign followed by a subpattern label), denoting a set of subpatterns, the set of subpatterns are tried in the order listed in the subpatterns file.

- For a type 4 token (a single asterisk "*"), there are two alternatives:
  1. the token is processed without consuming anything from the sentence
  2. one word is consumed from the remaining text.

  The first alternative is tried first. Alternative 2 is tried when the program backtracks to this token.

- For a type 5 token (three asterisks "***"), indicating that any number of words can be consumed from the sentence for this token. This token is first processed without consuming anything from the sentence. On backtracking, one word is consumed from the sentence. Each time the program backtracks to this token, an additional word is consumed from the sentence.

- A type 6 or type 7 token ("[1]" or "[2]") indicates that one or more words are to be consumed from the remaining text. The token is processed differently depending on whether the token occurs at the end of the pattern:
  1. In the case where the token occurs *before* the end of the pattern: One word is consumed from the remaining text when the token is first processed. If, however, the first item in the remaining text is a punctuation mark, enough items are consumed from the text so as to include one word that is not a punctuation mark. (The reason for this is that the words that match with this token will be identified as the cause or effect in a causal relation. Punctuation marks alone cannot represent the cause or effect.) On each subsequent backtracking, an additional word or punctuation mark is consumed from the sentence.
  2. In the case where the token is the last token of the pattern: All the remaining text is consumed if the remaining text includes at least one word that is not a punctuation mark. In other words, the token matches the rest of the words in the sentence.

- A type 8, type 10 or type 12 token ("[X:1]", "[X:2]" or "[X:0]") indicates a particular type of phrase. For example, "[N:0]" indicates a noun phrase. Ambiguity about what to consume from the text occurs when a noun phrase is embedded in another noun phrase, e.g.

  [N [N the book] in [N the store]]

> In such a case, the outermost phrase (i.e. "[N [N the book] in [N the store]]") is consumed when the token is first processed. On subsequent backtracking, the embedded phrase "[N the book]" is consumed instead.

If the pattern matching program fails to find the item indicated by a token, the program backtracks to the previous token processed. By "backtracking," I mean that the program replaces the token that it has just removed from the pattern and also replaces the last item removed from the text, and "reprocesses" the previous token. If this previous token indicates alternative items to search and these alternatives have not been tried before, the program looks for these alternative items at the beginning of the remaining text. If found, the item is consumed from the sentence and the program moves forward again to the next iterative step and processes the next token.

If, on the other hand, none of the alternative items are found or if the token does not specify alternative items, then the program continues to backtrack to the previous token processed. If the program finally backtracks to the beginning of the pattern and no alternative items are found for the first token, then the pattern has failed to match the sentence and the processing of this pattern terminates. From the above description, it is clear that a pattern is equivalent to a finite state transition network, and that each token corresponds to an arc in this network.

When a pattern succeeds in matching a sentence, two side-effects are carried out by the pattern matching program:

- the words that match with the slots (i.e. the words that were consumed when the slots were processed) are extracted as the cause or effect (depending on the type of slots).
- the words in the sentence (excluding punctuation marks) that match with tokens of type 2, 3 and 13 (i.e. tokens indicating parts-of-speech, verb groups or literal words) are flagged so that they are not allowed to match with any token of type 2, 3 and 13 in subsequent patterns. However, this rule is overridden when the "~" symbol is added to the end of a token. Words that match with a token with the "~" symbol are not flagged. Also, a token with the "~" symbol is allowed to match words that are flagged.

A sentence may contain more than one causal relation, and a pattern may be able to match more than one part of a sentence. For example, the pattern "[2] because [1]" occurs twice in the following sentence:

> The car didn't brake in time because the road was slippery, and the road was slippery because it hadn't been cleared of snow.

To ensure that every occurrence of a pattern in a sentence is found, the pattern matching program applies a pattern repeatedly to a sentence until the pattern fails. I have explained earlier that when a pattern is found to match a sentence, words that match with tokens of type 2, 3 and 13 (tokens indicating parts-of-speech, verb groups and literal words) are flagged. When the same pattern is applied the second time, the flagged words are not allowed to match with part-of-speech, verb-group and literal tokens in the pattern. This is why repeated application of the same pattern to a sentence can identify different instances of the causal relation.

This is also why a pattern must contain at least one token of type 2, 3 or 13. If a pattern does not contain a token of type 2, 3 or 13, no word is flagged when the pattern succeeds in matching a sentence. So, the pattern may be applied to a sentence repeatedly without ever failing, thus resulting in an infinite loop.

All the patterns in each set of patterns are applied to each sentence (or to each pair of adjacent sentences, in the case of the patterns that cross sentence boundaries). The patterns are applied in the order listed in the set of patterns. In practice, each pattern is indexed by a set of keywords. During pattern matching, a pattern is tried only if the sentence contains one of the keywords used to index the pattern.

C. Subpatterns That Are Frequently Used

To make it easier for the reader to understand the patterns listed in Section E, this section describes the purpose of some subpatterns that occur frequently in the patterns.

&C refers to a set of subpatterns that often introduce a clause. The subpatterns are used to find the beginning of a clause. The tokens *&C1*, *&C2*, *&C_*, *&C1_* and *&C2_* have a similar purpose. The subpatterns referred to by these tokens are variations of the set of subpatterns referred to by *&C*.

&S refers to a set of subpatterns that often occur at the beginning of sentence. *&S_* is a variation of *&S*.

&. refers to a set of subpatterns that often indicate the end of a clause (and this of course includes the period that marks the end of a sentence). *&._* is a variation of *&.* .

&AUX refers to a set of auxiliary verbs. *&AUX_* is a variation of *&AUX*.

I adopted the convention of adding the underscore character "_" to the end of a subpattern label if the set of subpatterns referred to includes the empty subpattern. For example, the token *&C* refers to the same set of subpatterns as the token *&C_* except that the set of subpatterns indicated by *&C_* includes an empty subpattern.

D. Verb-Group Labels Used In The patterns

Only three verb-group labels are used in the patterns listed in Sections E1 and E2 (which list patterns involving causal links):

$that refers to the group of verbs that are often used with the complementizer "that", for example:

acknowledge (that)
agree (that)
note (that)
say (that)

$through refers to the group of verbs that are often used with the preposition "through", for example:

browse (through)
channel (through)
crash (through)
sail (through)

$as refers to the group of verbs that are often used with the preposition "as", for example:

acknowledge (as)
employ (as)
identify (as)
portray (as)

The above three groups of verbs were identified using the Longman Dictionary of Contemporary

English (2nd ed.).

In the patterns listed in Section E3 (patterns involving causal verbs), most of the verb-group labels refer to groups of causal verbs. These groups are not the same as the verb categories listed in Appendix 3. The groups of causal verbs referred to by the verb group labels are defined by the prepositions that they expect. For example,

$+into refers to the group of causal verbs that are used with the preposition *into*, e.g.
bamboozle (into)
bludgeon (into)
cajole (into)
force (into)

$+out_of refers to the group of causal verbs that are used with the preposition *out of*, e.g.
bamboozle (out of)
coax (out of)
con (out of)
shame (out of)

$away refers to the group of causal verbs that are used with the particle *away*, e.g.
blow (away)
erode (away)
send (away)
sweep (away)

E. Linguistic Patterns

There are three sets of linguistic patterns. The patterns listed in Sections E1 and E2 involve "causal links" (described in Chapter 3, Section 3.2) and *if-then* conditionals, whereas the patterns listed in Section E3 involve causal verbs (described in Chapter 3, Section 3.3). The patterns in Sections E1 and E3 are applied to each sentence at a time. They are used to identify cause and effect within a sentence. The patterns listed in Section E2 are applied to each pair of adjacent sentences. They are used to identify instances where the cause and effect occur in adjacent sentences.

In the three lists of patterns, the symbol "#" indicates a comment. Words following the "#" sign are comments. Some of the patterns are preceded by the "#" symbol, indicating that these patterns were not found to be effective with the Wall Street Journal text used in this study. They may however be effective with other document collections.

In the lists, the first column indicates the type of relation that the pattern is supposed to identify. "C" in the first column indicates that the pattern is for identifying the causal relation. "-" in the first column indicates a null relation, i.e. the pattern does not indicate any relation. The patterns assigned a null relation are used for their side effects. The words that match with these patterns are flagged so that they may not match with part-of-speech, verb-group and literal tokens (tokens of type 2, 3 and 13) in subsequent patterns.

E1. Patterns involving causal links and *if-then* conditionals that link two phrases within a sentence

Relation Pattern

# The following set of patterns indicate causal relations across two sentences.

```
# They are listed here to preclude them being used for single sentences.
 -      &S so
 -      &S &THIS &AUX &ADV_ &ADJ &EFFECT/RESULT of
 -      &S &THIS &AUX &ADV_ due &RB_ to
 -      &S &THIS &AUX &ADV_ owing to
 -      &S &THIS &AUX &ADV_ &ADJ &REASON why
 -      &S &THIS &AUX &ADV_ why
 -      &S &THIS &AUX &ADV_ because
 -      &S &THIS &AUX &ADV_ on account of
 -      &S it &AUX &ADV_ for &THIS &REASON that
 -      there *** &REASON &FOR/WHY
 -      there *** &EFFECT/RESULT &OF/FROM
 -      there *** &EFFECT/RESULT to
 -      &THIS &AUX *** not *** only &REASON &FOR/WHY
 -      &THIS &AUX *** not *** only &EFFECT/RESULT &OF/FROM
 -      &THIS &AUX *** not *** only &EFFECT/RESULT to
 -      there &AUX *** &ANOTHER &REASON &FOR/WHY
 -      there &AUX *** &ANOTHER &EFFECT/RESULT &OF/FROM
 -      there &AUX *** &ANOTHER &EFFECT/RESULT to
 -      &S &ADJ &EFFECT/RESULT :
 -      &S &ADJ &EFFECT/RESULT &AUX
 -      &S &ADJ &REASON :
 -      &S &ADJ &REASON &AUX
 -      &S &ADJ &CAUSE :
 -      &S &ADJ &CAUSE &AUX
 -      &S &THIS &AUX &ADV_ &CRITICAL in
 -      &S &THIS &AUX &ADV_ &NECESSARY &TO/FOR

 C      how &CRITICAL [1] &AUX to the success of [2] &._

 -      &NO &EFFECT/RESULT

 C      &C1_ [1] &AND &THIS/IT &AUX_ &ADV_ &MAKE it possible for [2] &._
 -      as possible
 C      &C1_ [1] &AND &THIS/IT &AUX_ &ADV_ &MAKE [2] possible
 C      &C1_ &[1](AND_THIS) &AUX_ &ADV_ &MAKE it possible for [2] &._
 C      &C2_ [2] &AND [1] &AUX_ &ADV_ &MAKE &THIS/IT possible
 C      &C1_ &[1](AND_THIS) &AUX_ &ADV_ &MAKE [2] possible

# C     &C1_ [1] &WHICH [1] &REPORT &EFFECT/RESULT in [2] &._
# C     &C1_ [1] &WHICH [1] &REPORT &EFFECT/RESULT &OF/FOR [2] &._

 -      &PERSON &AUX_ &ADV_ fulfill* *** dream of
 C      &C1_ &[1](AND_THIS) &AUX_ &ADV_ fulfill* *** dream of [2] &._

 C      [1] &AND &THEREFORE [2] &._
 C      [1] and in consequence [2] &._
 C      &C1_ [1] so that [2] &.

 -      as *** , so &BE
 -      as *** , so &HAVE
 -      or so
 -      if so
 -      &NOT so
 -      more so
 -      so &BE
 -      so &HAVE
 -      &THINK so

 C      &C1_ [1] so [1] that [2] &._
 C      &C1_ [1] so [1] as to [2] &._
 C      &C2_ [2] so [j:1] &AUX [1] &._
# C    &C1_ [1] too [1] to [2] &._

 -      so @RB
 -      so *** as
 -      &BE so
 -      so @J
 -      so @V
 -      so @W
 C      &C1_ [1] so [2] &._

 C      [1] &AND for &THIS &REASON [2] &._

 -      &NOT as &DT &EFFECT/RESULT
 C      [1] &AND as &ADJ &EFFECT/RESULT of &THIS[1],[2] &._
 C      [1] &AND as &ADJ &EFFECT/RESULT [2] &._

 -      &NOT on account
 C      [1] &AND on &THIS account [2] &._
 C      [1] &AND on account of &THIS[1],[2] &._
 C      [1] &AND because of &THIS[1],[2] &._
 C      [1] &AND on &THIS &GROUND [2] &._
 C      [2] &AND owing to &THIS[1],[2] &._

 C      &C2_ [2] for &ADJ &REASON *** : [1]
 C      &C2_ [2] on the &GROUND that [1] &._

 C      &C2_ [2] for &ADJ &REASON that [1] &._
 C      &C2_ [2] on &ADJ &GROUND *** : [1]
 C      &C1_ [1] with the &EFFECT/RESULT that [2] &._

 -      &NOT because
 C      it &AUX &ADV_ because of [1] that [2] &._
 C      it &AUX &ADV_ because [1] that [2] &._
 C      it &AUX &ADV_ due &RB_ to [1] that [2] &._

 C      &C2_ &[2](AND_THIS) &AUX &ADV_ because of [1] &._
 C      &C2_ &[2](AND_THIS) &AUX &ADV_ because [1] &._
 C      &C2_ &[2](AND_THIS) &AUX &ADV_ on account of [1] &._

 C      &C because of &[N:1],[2]
 C      &C2_ [2] because of [1] &._

 -      &NOT owing to
 C      &C owing to &[N:1],[2]
 C      &C2_ [2] owing to [1] &._
```

```
-       &NOT on the &GROUND
C       &C on the &GROUND of &[N:1],[2]
C       &C2_ [2] on the &GROUND of [1] &._
-       on &MANY &GROUND
-       on &DT_ &SAME/DIFFERENT &GROUND
C       &C on [j:1] &GROUND [2] &._
C       &C2_ [2] on [j:1] &GROUND

C       &C on account of &[N:1],[2]
C       &C2_ [2] on account of [1] &._

-       &NOT for &REASON
C       &C for reasons of &[N:1],[2]
C       &C2_ [2] for reasons of [1] &._
-       for &MANY reasons
-       for &DT_ &SAME/DIFFERENT &REASON
C       &C for [j:1] reasons [2] &._
C       &C2_ [2] for [j:1] reasons

-       &NOT as &DT &EFFECT/RESULT
C       &C as &ADJ &EFFECT/RESULT of &[N:1],[2]
C       &C2_ [2] as &ADJ &EFFECT/RESULT of [1] &._

-       &NOT for the sake
C       &C for the sake of &[1], [2] &._
C       &C for the sake of [1] &[C:2].
C       &C for the sake of [G:1] [2] &._
C       &C2_ [2] for the sake of &[G:1].
C       &C2_ [2] for [1] sake

-       &NOT by reason
C       &C by reason of &[N:1],[2]
C       &C2_ [2] by reason of [1] &._

C       with the &ADVENT of &[N:1],[2]

-       &NOT due &RB_ to
-       due to expire
C       &C due &RB_ to &[N:1],[2]
C       &C2_ &[2](AND_THIS) &AUX &ADV_ due &RB_ to [1] &._
C       &C2_ [2] due &RB_ to &[N:1]
C       &C2_ [2] due &RB_ to [1] &._
C       &C on account of &[N:1],[2]
C       &C2_ [2] on account of [1] &._

C       &C because &[C:1],[2]
C       [C:2] only because &[C:1].
C       [C:2] because &[C:1].
C       &C2_ [2] only because [1] &._
C       &C2_ [2] because [1] &._

-       &NOT for
C       [C:2] for &[C:1].
C       &C2_ [2] for &[C:1].
```

```
-       ever since
-       time since
-       @DT @J since
-       since *** &TIME
-       since &[N:0] ,
-       since &[N:0] &.
-       &TIME2 since
-       &S &[N:0] since
-       , &[N:0] since
-       &S &[N:0] [P:0] since
-       , &[N:0] [P:0] since
-       &HAVE since
-       since then
-       since before
-       since after
-       since *** , *** &HAVE &ADV_ &VBD
-       &HAVE &ADV_ &VBD *** since
C       &C since &[C:1],[2]
C       &C2_ [2] since [1] &._

-       as *** as
-       such *** as
-       same *** as
-       &AS_BEFORE as
-       as &AS_AFTER
-       @J as [N:0] &BE
-       @J as [N:0] seems
-       , as &[N:0] &RB_ $that &,/.
-       " as &[N:0] &RB_ $that &,/.
-       &C as
-       &C2_ [2] as &"_ &[N:0] &"_ &.
-       &C2_ [2] as &"_ &[N:0] &"_ [P:0] &.
-       $as *** as
C       &C2_ &[C:2] as &[N:1] [1] &._

# C     &C having &[1], [2] &._
# C     &C having [1] [C:2]
# C     &C2_ [2] having [1] &._
# C     having &[1], [2] &._
# C     having [1] [C:2]

-       &NOT &HAVE &ADJ &EFFECT
C       &C1_ &[1](AND_THIS) &AUX_ &ADV_ &HAVE &ADJ &EFFECT/RESULT *** : [2]
C       &C1_ &[1](AND_THIS) &AUX_ &ADV_ &HAVE &ADJ &EFFECT on [2] &._

-       &NOT &ADJ &EFFECT/RESULT
C       &C2_ &[2](AND_THIS) &AUX &ADV_ &ADJ &EFFECT/RESULT &OF/FROM [1] &._
C       &C1_ [1] &AND &ADJ &EFFECT/RESULT *** : [2]
-       &BE &ADV_ &VBD/VBG
-       &BE &NOT &ADV_ &VBD/VBG
-       &BE &ADJ &.
C       &C1_ [1] &AND &ADJ &EFFECT/RESULT of &THIS &AUX [2] &._
C       &C1_ [1] &AND &ADJ &EFFECT/RESULT of &THIS [1] &AUX [2] &._
```

```
C       &C1_ [1] &AND &ADJ &EFFECT/RESULT &AUX [2] &._
C       &EFFECT/RESULT of [1] &AUX &ADV_ on [2] &._
C       &EFFECT of [1] on [2] : [2]
C       &EFFECT of [1] on [2] &AUX &ADV_ &VBD/VBG
C       &EFFECT of [1] on [2] &AUX [2] &._
-       &EFFECT of &[conj_N:0],
C       &EFFECT of &[conj_N:1] on [2] &._
-       &EFFECT &[P:0], *** on
C       &EFFECT of [1] on [2] &._
C       &EFFECT/RESULT of [1] : [2]
C       &EFFECT/RESULT of [1] &AUX &ADV_ that [2] &._
C       &EFFECT/RESULT of [1] &AUX &NOT *** but [2] &._
-       &BE &NOT
C       &EFFECT/RESULT of &[conj_N:1] &AUX [2] &._
-       &EFFECT/RESULT &[P:0],
C       &EFFECT/RESULT of [1] &AUX &[not_cc:2].

C       &EFFECT [1] &HAVE on [2] &._
C       &EFFECT [1] &AUX_ &ADV_ &HAVE on [2] &._

-       &NOT &ADJ &CAUSE
C       &C1_ &[1](AND_THIS) &AUX &ADV_ &ADJ &CAUSE &OF/IN [2] &._
C       &C1_ &[2](AND_THIS) &AUX_ &ADV_ &HAVE &ADJ &CAUSE *** : [1]
C       &C2_ [2] &AND &ADJ &CAUSE *** : [1]
-       &BE &ADV_ &VBD/VBG
-       &BE &NOT &ADV_ &VBD/VBG
-       &BE &J &.
C       &C2_ [2] &AND &ADJ &CAUSE of &THIS/IT &AUX [1] &._
C       &C2_ [2] &AND &ADJ &CAUSE of &THIS [2] &AUX [1] &._
C       &C2_ [2] &AND &ADJ &CAUSE &AUX [1] &._
C       &CAUSE of [2] : [1]
C       &CAUSE of [2] &AUX &ADV_ that [1] &._
C       &CAUSE of [2] &AUX &ADV_ &NOT *** but [1] &._
-       &AUX &NOT
-       &CAUSE of &[conj_N:0],
C       &CAUSE of &[conj_N:2] &AUX [1] &._
-       &CAUSE &[P:0],
C       &CAUSE of [2] &AUX &[not_cc:1].

-       &HAVE &ADJ_ &REASON for
-       &NOT &ADJ &REASON for
-       &NOT &ADJ &REASON
-       there &AUX &ADJ_ &REASON
-       &REASON to @V
C       &C1_ &[1](AND_THIS) &AUX &ADV_ &ADJ_ &REASON [2] &._
-       &NOT why
C       &C1_ &[1](AND_THIS) &AUX &ADV_ why [2] &._
C       &C2_ [2] &AND &ADJ &REASON *** : [1]
-       &BE &ADV_ &VBD/VBG
-       &BE &NOT &ADV_ &VBD/VBG
-       &BE &J &.
C       &C2_ [2] &AND &ADJ &REASON &OF/FOR &THIS/IT &AUX [1] &._
C       &C2_ [2] &AND &ADJ &REASON &OF/FOR &THIS/IT [2] &AUX [1] &._
C       &C2_ [2] &AND &ADJ &REASON why &AUX [1] &._
```

```
C	&C2_ [2] &AND &ADJ &REASON why [2] &AUX &ADV_ &BECAUSE/THAT [1] &._
C	&C2_ [2] &AND &ADJ &REASON why [2] &[VP:2] &AUX [1] &._
C	&C2_ [2] &AND &ADJ &REASON &AUX [1] &._
C	&REASON &FOR/WHY [2] : [1]
C	&REASON &FOR/WHY [2] &AUX &ADV_ &BECAUSE/THAT [1] &._
C	&REASON [2] &AUX &ADV_ &BECAUSE/THAT [1] &._
C	&REASON &FOR/WHY [2] &AUX &ADV_ &NOT *** but [1] &._
-	&BE &NOT
-	&REASON &FOR &[conj_N:0],
C	&REASON &FOR &[conj_N:2] &AUX [1] &._
-	&REASON &[P:0],
C	&REASON &FOR [2] &AUX &[not_cc:1].
C	&REASON why [2] &[VP:2] &AUX &[not_cc:1].
C	&REASON &[C:2] &AUX &[not_cc:1] [1] &._
C	why [2] &AUX &ADV_ that [1] &._

C	&C1_ &[1](AND_THIS) &AUX_ &ADV_ &PROVIDE a basis for [2] &._
C	&C1_ &[1](AND_THIS) &AUX &ADV_ a basis for [2] &._

-	&NOT &CRITICAL &TO/FOR
-	&NOT &CRITICAL
-	important &TO/FOR &DT_ &PERSON
C	&TO/FOR [v:2] , it &AUX &ADV_ &CRITICAL to &[VP:1].
C	&TO/FOR [v:2] &[2], it &AUX &ADV_ &CRITICAL to &[VP:1].
C	it &AUX &ADV_ &CRITICAL &TO/FOR &[VP:1] to &[VP:2].
C	for the &[2], it &AUX &ADV_ &CRITICAL to &[VP:1].
C	it &AUX &ADV_ &CRITICAL &TO/FOR [2] to &[VP:1].
-	it &AUX &ADV_ &CRITICAL to
-	too &CRITICAL to @V
C	&C1_ &[1](AND_THIS) &AUX &ADV_ &CRITICAL &TO/FOR [2] than [1] &._
C	&C1_ &[1](AND_THIS) &AUX &ADV_ &CRITICAL &TO/FOR [2] &._
C	&C1_ &[1](AND_THIS) &AUX &ADV_ &CRITICAL in &[G:2].
C	more &CRITICAL &TO/FOR [2] than [1] &._
C	&CRITICAL &TO/FOR [2] &AUX &[not_cc:1].
C	&CRITICAL &TO/FOR &[2], [1] &._

-	&NOT &ADJ &FACTOR
C	&C1_ &[1](AND_THIS) &AUX_ &ADV_ &EMERGE as &ADJ &FACTOR &OF/IN [2]
	&._
C	&C1_ &[1](AND_THIS) &AUX &ADV_ &ADJ &FACTOR &OF/IN [2] &._
C	&FACTOR &OF/IN [2] &AUX &ADV_ &J *** : [1]
C	&FACTOR &OF/IN [2] &AUX &ADV_ that [1] &._
-	&BE &ADV_ &VBD/VBG
-	&BE &ADV_ &NOT &VBD/VBG
-	&BE &ADV_ &J &.
C	&FACTOR &OF/IN [2] &AUX &ADV_ &NOT *** but [1] &._
-	&FACTOR &OF/IN &[conj_N:0],
C	&FACTOR &OF/IN &[conj_N:2] &AUX [1] &._
-	&FACTOR &[P:0],
-	&BE &NOT
C	&FACTOR &OF/IN [2] &AUX &[not_cc:1].

C	&C1_ &[1](AND_THIS) &AUX &ADV_ initiator* of [2] &._
```

```
 -      &NOT &ADJ_ &OBSTACLE to
 C      &C1_ &[1](AND_THIS) &AUX &ADV_ &ADJ_ &OBSTACLE to [2] &._

 C      &C1_ &[1](AND_THIS) &AUX_ &ADV_ &HAVE the potential for [2] &._

 C      &C1_ &[1](AND_THIS) &AUX &ADV_ &ADJ_ predictor* of [2] than [2] &._
 C      &C1_ &[1](AND_THIS) &AUX &ADV_ &ADJ_ predictor* of [2] &._

 C      &C1_ [1] &AUX &ADV_ held responsible for [2] &._
 C      &HOLD [1] responsible for [2] &._
 C      &C1_ [1] responsible for [2] &._

# -     &TAKE *** advantage
# -     &C &THIS/IT &AUX_ &ADV_ &HAVE &ADJ_ advantage*
# -     &C @IN advantage* ,
# C     &C1 &[1](AND_THIS) &AUX_ &ADV_ &HAVE &ADJ_ advantage of [2] &._
# C     advantage* * &OF/IN [1] &TO/FOR [2] : [2]
# C     advantage* * &OF/IN [1] over *** : [2]
# C     advantage* * &OF/IN [1] : [2]
# -     advantage* in &[N:0]
# -     advantage* * &OF/IN &[conj_N:1],
# C     advantage* * &OF/IN &[conj_N:1] &AUX &[not_cc:2].
# -     advantage* &[P:0],
# -     advantage* * &OF/IN [G:0] ,
# -     advantage* * &OF/IN [G:0] [P:0] ,
# C     advantage* * &OF/IN [1] &TO/FOR [2] &BE &ADV_ &VBD/VBG
# C     advantage* * &OF/IN [1] &TO/FOR [2] &BE [2] &._
# -     advantage* * &OF/IN *** over *** &BE &ADV_ &VBD/VBG
# C     advantage* * &OF/IN [1] over *** &AUX [2] &._
# -     advantage* * &OF/IN *** &BE &ADV_ &VBD/VBG
# C     advantage* * &OF/IN [1] &AUX &[not_cc:2].
# -     &C &THIS/IT &AUX_ &ADV_ &HAVE &ADJ_ disadvantage*
# -     &C it &AUX_ &ADV_ &HAVE &ADJ_ disadvantage*
# C     &C1 &[1](AND_THIS) &HAVE &ADJ_ advantage of [2] &._
# C     disadvantage* * &OF/IN [1] &TO/FOR [2] : [2]
# C     disadvantage* * &OF/IN [1] over *** : [2]
# C     disadvantage* * &OF/IN [1] : [2]
# -     disadvantage* in &[N:0]
# -     disadvantage* * &OF/IN &[conj_N:1],
# C     disadvantage* * &OF/IN &[conj_N:1] &AUX [2] &._
# -     disadvantage* &[P:0],
# -     disadvantage* * &OF/IN [G:0] ,
# -     disadvantage* * &OF/IN [G:0] [P:0] ,
# C     disadvantage* * &OF/IN [1] &TO/FOR [2] &BE &ADV_ &VBD/VBG
# C     disadvantage* * &OF/IN [1] &TO/FOR [2] &BE [2] &._
# -     disadvantage* * &OF/IN *** over *** &BE &ADV_ &VBD/VBG
# C     disadvantage* * &OF/IN [1] over *** &AUX [2] &._
# -     disadvantage* * &OF/IN *** &BE &ADV_ &VBD/VBG
# C     disadvantage* * &OF/IN [1] &AUX &[not_cc:2].

 -      downside to *** &AUX &ADV_ &VBD/VBG
 C      downside to [1] &AUX &[not_cc:2].

 -      if not~
 -    &KNOW if
 -    even if
 -    if *** , &WH
 -    if any &,/.
 -    if @RB &,/.
 -    if *** &NOT have to
 C    &IF [2] &WANT to [2] , [1] &MUST [1]
 -    if *** &MUST
 C    &IF [2] &WANT to [2] , it &AUX &ADV_ &CRITICAL [1]
# C   &IF [2] , [1] &MUST [1] &._
# C   &IF [2] &[N:1] &MUST [1] &._
# C   &IF [2] it &AUX &ADV_ &CRITICAL [1]

 C    &C2_ &[2](AND_THIS) &AUX possible only &IF [1] &._
 C    &C2_ &[2](AND_THIS) &AUX only possible &IF [1] &._
 C    &C2_ &[2](AND_THIS) &AUX possible &IF [1] &._
 C    &C2_ [2] only &IF [1] &._
 C    only &IF [1] can [2] &._
 C    &C1_ [1] &IF [2] &BE to [2] &._

 C    [2] &WILL [2] , [N:0] $that , &IF [1] &._
 C    [2] &WILL [2] , $that [N:0] , &IF [1] &._
 C    &C/, &IF [1] then [2] &._
 C    &C/, &IF &[1], [2] &WILL [2] &._
 C    &C/, &IF [1] &[N:2] &WILL [2] &._
 C    &C2_ [2] &WILL [2] &IF [1] &._
 C    &IF [1] then [2] &._
 C    &IF &[1], [2] &WILL [2] &._
 C    &IF [1] &[N:2] &WILL [2] &._

 C    &ATTRIBUTE [2] to [1] &._
 C    &C2_ &[2](AND_THIS) &AUX &ADV_ &ATTRIBUTE to [1] &._

 C    &BLAME [1] for [2] &._
 C    &C1_ &[1](AND_THIS) &AUX &ADV_ blame* for [2] &._

 -    &NOT &NECESSARY &TO/FOR
 C    &C1_ &TO/FOR [2] it &AUX &ADV_ &NECESSARY &TO/FOR [2] &._
 C    &C1_ &[1](AND_THIS) &AUX &ADV_ &NECESSARY &TO/FOR [2] &._

# -   &REQUIRE by
# -   &REQUIRE that
# -   &REQUIRE to
# -   &REQUIRE &[N:0] to
# -   &[N:0] [G:0] &REQUIRE
# -   &NOT &REQUIRE
# C   &C2_ &[2](AND_THIS) &AUX_ &ADV_ &REQUIRE [1] &._
# C   for [2] &AUX_ &ADV_ &REQUIRE [1] &._
# C   [I:2] &AUX_ &ADV_ &REQUIRE [1] &._
# C   [I:2] , it &AUX_ &ADV_ &REQUIRE [1] &._
# C   [G:2] &AUX_ &ADV_ &REQUIRE [1] &._

 -    &NOT &PLAY
 C    &C1_ &[1](AND_THIS) &AUX_ &ADV_ &PLAY &ADJ role in [2] &._
 -      role of *** &AUX &ADV_ &VBD/VBG
 C      role of [1] &AUX &NOT *** but [2] &._
 -      role of [1] &AUX &NOT
 C      role of [1] &AUX &[not_cc:2].

 -      &NOT &DELIVER
 C      &C1_ &[1](AND_THIS) &AUX_ &ADV_ &DELIVER &ADJ_ benefit* to [2] &._

 C      &C1_ &[1](AND_THIS) &STAND in the way of [2] &._

 -      &NOT sufficient to
 C      &C1_ &[1](AND_THIS) &AUX_ &ADV_ sufficient to [2] &._

 C      &C2_ [2] in the wake of [1] &._
 C      &C2_ [2] in the aftermath of [1] &._
 C      &C2_ [2] amid [1] &._

 C      &C1_ &[1](AND_THIS) &AUX &ADV_ &CRITICAL in &[G:2].

 C      it &AUX &ADV_ because of [1] that [2] &._

 -      &NOT for
 C      &FOR &[G:2] by &[G:1].

 -      &VERB_BY *** by
 C      &C by [vbg:1] , [2] &._
 C      &C by [vbg:1] &[1], [2] &._
 C      &C by &[G:1] &[C:2].
 C      &C by [vbg:1] [2] &._
 C      &C2_ [I:2] by &[G:1].
 C      &C2_ [2] by &[G:1].

 -      exchange for
 -      @J for
 -      &HAVE &[N:0] for
 -      &AUX &[N:0] for
 -      &,/. &[N:0] for
 -      &,/. &ADJ_ @N for
 -      &C &[N:0] for
 -      &C &ADJ_ @N for
 -      &NOUN_FOR *** for
 -      &VERB_FOR *** for
 C      &C for [G:1] , [2] &.
 C      &C for [G:1] &[1], [2] &._
 C      &C for &[G:1] &[C:2].
 C      &C for [vbg:1] [2] &._
 C      &[N:2] [I:2] for &[G:1].
 C      [I:2] for &[G:1].
 C      &C2 &[VP:2] for &[G:1].
 C      &C2_ [2] &[VP:2] for &[G:1].

 -      &NOT through
# -     through @IN
# -     through &DT_ &PERSON
# -     through it
# -     through &TIME
# -     $through *** &THROUGH/TO
# C     &C &ADV_ through &[1], [2] &._
# C     &C &ADV_ through [1] [I:2] &.
# C     &C &ADV_ through [1] [I:2] [2] &._
# C     &C &ADV_ through [1] &[C:2].
# C     &C &ADV_ through &[N:1] [2] &._
# C     &C2_ &[2], through &[1], [2] &._
# C     &C2_ [2] through &[G:1].
# C     [I:2] through [1] &._
# C     &C2_ [2] through [1] &._
# C     &ACHIEVE [2] through [1] &._

  C     &C through [vbg:1] , [2] &._
  C     &C through [vbg:1] &[1], [2] &._
  C     &C through &[G:1] &[C:2].
  C     &C through [vbg:1] [2] &._
  C     &C2_ [I:2] through &[G:1].
  C     &C2_ [2] through &[G:1].

  C     &C in order &TO/FOR &[2], [1] &._
  C     &C1_ [1] in order &TO/FOR [2] &._
  C     &C in order that &[2], [1] &._
  C     &C1_ [1] in order that [2] &._

  C     &C1_ [1] in an effort to [2] &._

# -     to @VBG
# -     &NOT to
# -     &AUX to
# -     &AUX &[N:0] to
# -     &AUX for &[N:0] to @V
# -     &AUX &ADV_ &J to @V
# -     &AUX &PERSON to @V
# -     @VBG &TO
# -     @J to
# -     &HAVE &TO
# -     take *** time to
# -     @W &ADV_ to
# -     &TAKE &TIME to @V
# -     &NOUN_TO *** &TO
# -     $to *** &TO
# C     &C1_ [1] &[VP:1] to &ADV_ &[VP:2]
# -     &[N:0] to &ADV_ @V

# Sentence initial infinitive phrases
  -     to @VBG
  -     &NOT to
  -     to be sure
  -     agree* *** to
  C     &S &"_ to &ADV_ [v:2] , [1] &._
  C     &S &"_ to &ADV_ [v:2] &[2], [1] &._
  C     &S &"_ to &ADV_ &[VP:2] &[C:1].
```

```
 C      &C2_ &[2](AND_THIS) &AUX &ADV_ &NOT dependent on *** but on [1] &._
 -      &NOT dependent
 C      &C2_ &[2](AND_THIS) &AUX &ADV_ dependent on [1] &._

# -     &NOT being
# C     &C being &[1], [2] &._
# C     &C being [1] &[C:2].
# C     &C2_ [2] , being [1] &._

 -      . * [P:0] , &ADJ &EFFECT/RESULT of
 C      &C2_ [2] , &ADJ &EFFECT/RESULT of [1] &,/.
 C      &C2_ [2] -- &ADJ &EFFECT/RESULT of [1] &,/.

 -      &NOT in response
 C      &C in response to &[N:1],[2]
 C      &C2_ [2] in response to [1] &._

 C      &C1_ [1] &AUX credited with [2] &._
```

E2. Patterns involving causal links that link two sentences

Note: The token "." in the patterns indicates the end of a sentence.

Relation	Pattern
C	[1] . &S_ &THEREFORE [2]
C	[1] . &S_ in consequence [2]
C	[1] . &S_ and in consequence [2]
C	[1] . &S_ it &AUX &THEREFORE [2]
-	as *** , so &BE
-	as *** , so &HAVE
-	so as
-	or so
-	if so
-	&NOT so
-	more so
-	so @J
-	so @RB
-	so &BE
-	so &HAVE
-	so @V
-	&THINK so
-	so @W
-	so far
-	so|RB-called*
C	[1] . &S_ so [2]
-	&NOT because
C	[1] . &S_ because of &THIS,[2]
C	[1] . &S_ owing to &THIS,[2]
C	[1] . &S_ for &THIS &REASON [2]
C	[1] . &S_ as &ADJ &EFFECT/RESULT of &THIS,[2]
C	[1] . &S_ as &ADJ &EFFECT/RESULT [2]
C	[1] . &S_ on &THIS account [2]
C	[1] . &S_ on account of &THIS,[2]
C	[1] . &S_ because of &THIS,[2]
C	[1] . &S_ on &THIS &GROUND [2]
C	[2] . &S_ owing to &THIS,[2]
C	[2] for &THIS &J_ &REASON . &S_ [1]
C	[2] for the following &J_ &REASON . &S_ [1]
C	[2] on &THIS &GROUND . &S_ [1]
C	[2] on the following &GROUND . &S_ [1]
C	&THIS &AUX *** &NOT *** only &EFFECT/RESULT &OF/FROM [1] . &S_ [2]
-	&NOT &ADJ &EFFECT/RESULT
-	no &EFFECT/RESULT
C	[1] . &S_ &THIS/HERE &AUX &ADV_ &ADJ &EFFECT/RESULT : [2]
C	[2] . &S_ &THIS &AUX &ADV_ &ADJ &EFFECT/RESULT of [1]

```
 -      &NOT due to
 C      [2] . &S_ &THIS &AUX &ADV_ due &RB_ to [1]
 -      &NOT owing to
 C      [2] . &S_ &THIS &AUX &ADV_ owing to [1]

 C      &THIS &AUX *** not *** only &REASON &FOR/WHY [2] . &S_ [1]
 -      &NOT &ADJ &REASON
 -      no &REASON
 -      &NOT why
 C      [2] . &S_ &THIS/HERE &AUX &ADV_ &ADJ &REASON why : [1]
 C      [2] . &S_ &THIS/HERE &AUX &ADV_ &ADJ &REASON : [1]
 C      [2] . &S_ &THIS/HERE &AUX &ADV_ why : [1]
 C      [1] . &S_ &THIS &AUX &ADV_ &ADJ &REASON why [2]
 C      [1] . &S_ &THIS &AUX &ADV_ why [2]
 C      [2] . &S_ &THIS &AUX &ADV_ because of [1]
 C      [2] . &S_ &THIS &AUX &ADV_ because [1]
 C      [2] . &S_ &THIS &AUX &ADV_ on account of [1]

 -      &NOT &ADJ &CAUSE
 C      [2] . &S_ &THIS/HERE &AUX &ADV_ &ADJ &CAUSE : [1]
 C      [1] . &S_ &THIS &AUX &ADV_ &ADJ &CAUSE of [2]

 C      [1] . &S_ it &AUX &ADV_ for &THIS &REASON that [2]
 C      [1] . &S_ it &AUX &ADV_ because of &THIS *** that [2]
 C      [1] . &S_ it &AUX &ADV_ due &RB_ to &THIS *** that [2]

# -     &REASON &FOR/WHY *** :
# -     &REASON &FOR/WHY *** &BE
# C     &ADJ_ @J &REASON &FOR/WHY [2] . &S_ [1]
# C     &MANY &REASON &FOR/WHY [2] . &S_ [1]

# -     &EFFECT/RESULT *** :
# -     &EFFECT/RESULT &OF/FROM *** $BE
# C     &ADJ_ @J &EFFECT/RESULT &OF/FROM [1] . &S_ [2]
# C     &MANY &EFFECT/RESULT &OF/FROM [1] . &S_ [2]

 C      &THERE &AUX &ADJ_ &REASON &FOR/WHY [2] . &S_ [1]
 C      &THERE &AUX &ADJ_ &EFFECT/RESULT &OF/FROM/TO [1] . &S_ [2]
 C      &THERE &AUX *** &ANOTHER &REASON &FOR/WHY [2] . &S_ [1]
 C      &THERE &AUX *** &ANOTHER &EFFECT/RESULT &OF/FROM/TO [1] . &S_ [2]

# C     [1] &HAVE &ADJ &EFFECT/RESULT . &S_ [2]
# C     [1] &HAVE &MANY &EFFECT/RESULT . &S_ [2]

 C      [1] . &S_ &DT &ADJ_ &EFFECT/RESULT of &THIS *** : [2]
 C      [1] . &S_ &DT &ADJ_ &EFFECT/RESULT &AUX *** : [2]
 C      [1] . &S_ &DT &ADJ_ &EFFECT/RESULT of &THIS *** &AUX &ADV_ that [2]
 -      &EFFECT/RESULT of *** &NOT &ADJ .
 C      &DT &ADJ_ &EFFECT/RESULT of [1] &AUX &ADJ . &S_ [2]
 -      &EFFECT/RESULT of &THIS *** &AUX &ADV_ &VBD
 C      [1] . &S_ &DT &ADJ_ &EFFECT/RESULT of &THIS *** &AUX [2]
 C      [1] . &S_ &DT &ADJ_ &EFFECT/RESULT : [2]
 C      [1] . &S_ &DT &ADJ_ &EFFECT/RESULT &AUX [2]
```

```
C       [2] . &S_ &DT &ADJ_ &REASON for &THIS *** : [1]
C       [2] . &S_ &DT &ADJ_ &REASON &AUX *** : [1]
C       [2] . &S_ &DT &ADJ_ &REASON for &THIS *** &AUX &ADV_ that [1]
-       &REASON for *** &NOT &ADJ .
C       &DT &ADJ_ &REASON for [2] &AUX &ADJ . &S_ [1]
-       &REASON for &THIS *** &AUX &ADV_ &VBD
C       [2] . &S_ &DT &ADJ_ &REASON for &THIS *** &AUX [1]
C       [2] . &S_ &DT &ADJ_ &REASON : [1]
C       [2] . &S_ &DT &ADJ_ &REASON &AUX [1]

C       [2] . &S_ &DT &ADJ_ &CAUSE of &THIS *** : [1]
C       [2] . &S_ &DT &ADJ_ &CAUSE &AUX *** : [1]
C       [2] . &S_ &DT &ADJ_ &CAUSE of &THIS *** &AUX &ADV_ that [1]
-       &CAUSE of *** &NOT &ADJ .
C       &DT &ADJ_ &CAUSE of [2] &AUX &ADJ . &S_ [1]
-       &CAUSE of &THIS *** &AUX &ADV_ &VBD
C       [2] . &S_ &DT &ADJ_ &CAUSE of &THIS *** &AUX [1]
C       [2] . &S_ &DT &ADJ_ &CAUSE : [1]
C       [2] . &S_ &DT &ADJ_ &CAUSE &AUX [1]

C       [1] . &S_ &THIS/IT *** &HAVE *** advantage* *** : [2]
C       [1] . &S_ &THIS/IT *** &HAVE *** advantage* of [2] &._
C       [1] . &S_ @IN advantage* , [2] &._
C       [1] . &S_ &DT &ADJ_ advantage* of &THIS *** : [2] &.
C       [1] . &S_ &DT &ADJ_ advantage* of &THIS *** &AUX [2] &.
C       [1] . &S_ &DT &ADJ_ advantage* &AUX [2] &.

C       [1] . &S_ &DT &ADJ_ disadvantage* of &THIS *** &AUX [2] &.
C       [1] . &S_ &DT &ADJ_ disadvantage* &AUX [2] &.

-       &NOT &ADV_ &CRITICAL
C       [1] . &S_ &THIS &AUX &ADV_ &CRITICAL in [vbg:2] [2] &._

-       &NOT &ADV_ &NECESSARY
C       [1] . &S_ &THIS &AUX &ADV_ &NECESSARY &TO/FOR [2] &._

C       [1] . &S_ &THIS &MAKE it possible for [2] &._
C       [1] . &S_ &THIS &MAKE [2] possible
C       [2] . &S_ [1] &MAKE &THIS possible &.
```

E3. Patterns involving causal verbs

Relation	Pattern
# EXCEPTIONS	
-	&NOT $
-	$ &TIME
-	&TIME &AUX_ &RB_ $
-	by &TIME
# "NP VERB NP ADJ" PATTERN	
C	&[subj_N:2] &AUX_ &VBD [j:2] &,_ &J_ by &[N:1]
-	[N:0] such as
C	&[subj_N:1] &AUX_ &RB_ @V &[N:2] &RB_ [j:2] &MISC5
# PASSIVE CONSTRUCTIONS	
C	&[subj_N:2] &AUX_ &RB_ &past_v+against against~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+among among~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+around around~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+as as~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+at at~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+back_on back~ on~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+back_to back~ to~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+between between~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+for for~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+from_v-ing from~ &[G:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+from from~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+in_favor_of in~ favor~ of~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+in_for in~ for~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+in_on in~ on~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+on on~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_von on~ by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+in in~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_vin in~ by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_vphr_in in~ by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+into into~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+of of~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+off_to off~ to~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+off_with off~ with~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_vdown_on down~ on~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+onto onto~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+out_in out~ in~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+out_of out~ of~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+out out~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+over_to over~ to~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+over over~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+round round~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+through through~ &[N:2] by &[N:1]
C	&[subj_N:2] &AUX_ &RB_ &past_v+throughout throughout~ &[N:2] by &[N:1]

```
 C      &[subj_N:2] &AUX_ &RB_ &past_v+to-v to~ &RB_ &[VP:2] by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_v+to to~ &[N:2] by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_v+towards towards~ &[N:2] by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_v+upon upon~ &[N:2] by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_v+with with~ &[N:2] by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vabout about~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vapart apart~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vaway away~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vback back~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vdown down~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vforth forth~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_voff off~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vopen open~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vout out~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vover over~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vround round~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vthrough through~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vtogether together~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vup up~ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vphr_away_with away~ with~ by &[N:1]
 C      &[subj_N:1] &AUX_ &RB_ &past_vphr_from from~ by &[N:2]
 C      &[subj_N:2] &AUX_ &RB_ &past_vphr_out+to out~ to~ &[N:2] by &[N:1]
 -      &past_vparticle @IN by &TIME
 C      &[subj_N:2] &AUX_ &RB_ &past_v+n &"_ &[N:2] &"_ by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_v+prep &[P:2] by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_v+v-ing &[G:2] by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_v+v &RB_ to~ &[VP:2] by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_vparticle @RP by &[N:1]
 C      &[subj_N:2] &AUX_ &RB_ &past_v_ by &[N:1]
# C     &[subj_N:2] &AUX_ &RB_ &past_v+v-ed &RB_ &[VPpast:2] by &[N:1]


# ACTIVE CONSTRUCTIONS

 -      &BE &VBD
 C      &[subj_N:1] &AUX_ &RB_ $+against &[N:2] against~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+among &[N:2] among~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+around &[N:2] around~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+as &[N:2] as~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+at &[N:2] at~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+back_on &[N:2] back~ on~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+back_to &MISC31
 C      &[subj_N:1] &AUX_ &RB_ $+between &[N:2] between~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+for &[N:2] for~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+from_v-ing &[N:2] from~ &[G:2].
 C      &[subj_N:1] &AUX_ &RB_ $+from &[N:2] from~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+in_favor_of &[N:2] in~ favor~ of~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+in_for &[N:2] in~ for~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+in_on &[N:2] in~ on~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $+on &[N:2] on~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $on &MISC18
 C      &[subj_N:1] &AUX_ &RB_ $+in &[N:2] in~ &[N:2]
 C      &[subj_N:1] &AUX_ &RB_ $in &MISC20
```

```
C       &[subj_N:1] &AUX_ &RB_ $phr_in in~ [2] &._
C       &[subj_N:1] &AUX_ &RB_ $+into &[N:2] into~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+of &[N:2] of~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+off_to &[N:2] off~ to~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+off_with &[N:2] off~ with~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $down_on &MISC28
C       &[subj_N:1] &AUX_ &RB_ $+onto &[N:2] onto~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+out_in &[N:2] out~ in~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+out_of &[N:2] out~ of~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+out &[N:2] out~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+over_to &[N:2] over~ to~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+over &[N:2] over~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+round &[N:2] round~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+that &MISC27
C       &[subj_N:1] &AUX_ &RB_ $+through &[N:2] through~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+throughout &[N:2] throughout~ &[N:2]
C       &C1_ [1] to &RB_ $+to-v &[N:2] to~ &RB_ &[VP:2].
C       &[subj_N:1] &AUX_ &RB_ $+to-v &[N:2] to~ &RB_ &[VP:2].
C       &[subj_N:1] &AUX_ &RB_ $+to &[N:2] to~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+towards &[N:2] towards~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+upon &[N:2] upon~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+with &[N:2] with~ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $about &MISC26
C       &[subj_N:1] &AUX_ &RB_ $apart &MISC25
C       &[subj_N:1] &AUX_ &RB_ $away &MISC24
C       &[subj_N:1] &AUX_ &RB_ $back &MISC23
C       &[subj_N:1] &AUX_ &RB_ $down &MISC22
C       &[subj_N:1] &AUX_ &RB_ $forth &MISC21
C       &[subj_N:1] &AUX_ &RB_ $off &MISC19
C       &[subj_N:1] &AUX_ &RB_ $open &MISC17
C       &[subj_N:1] &AUX_ &RB_ $out &MISC16
C       &[subj_N:1] &AUX_ &RB_ $over &MISC15
C       &[subj_N:1] &AUX_ &RB_ $round &MISC14
C       &[subj_N:1] &AUX_ &RB_ $through &MISC13
C       &[subj_N:1] &AUX_ &RB_ $together &MISC12
C       &[subj_N:1] &AUX_ &RB_ $up &MISC11
C       &[subj_N:1] &AUX_ &RB_ $phr_away_with away~ with~ &[N:2]
C       &[subj_N:2] &AUX_ &RB_ $phr_from from~ [1] &._
C       &[subj_N:1] &AUX_ &RB_ $phr_out+to &MISC10
C       &C1_ [1] &AUX_ &RB_ $phr_to to [2] &._
C       &C1_ [1] &AUX_ &RB_ $phr_up up [2] &._
C       &[subj_N:1] &AUX_ &RB_ $+n &[N:2] &"_ &[N:2]
C       &[subj_N:1] &AUX_ &RB_ $+prep &[N:2] &[P:2]
C       &[subj_N:1] &AUX_ &RB_ $+v-ing &[N:2] &[G:2]
C       &[subj_N:1] &AUX_ &RB_ $+v-ed &[N:2] &RB_ &[VPpast:2]
-       $+v &[N:0] &RB_ &VBD/VBG
C       &[subj_N:1] &AUX_ &RB_ $+v &[N:2] &RB_ &[VP:2].
C       &C1_ [1] [1] to~ &RB_ &V_ $+v &[N:2] &RB_ &[VP:2].
C       &[subj_N:1] &AUX_ &RB_ $particle &MISC40

-       $ @CD ,
-       $ @CD &.

C       &C1_ [1] , $_|VBG &[obj_N:2]
```

```
C       &[subj_N:1] &AUX_ &RB_ $_ &[obj_N:2]
C       &[subj_N:1] @MD to $_ &[obj_N:2]
-       &[subj_N:1] &AUX_ &RB_ @V &[N:0]_ by $_|VBG &[obj_N:2]
-       too @J
C       &[subj_N:1] &AUX &RB_ &J &MISC9
C       &[subj_N:1] &AUX &RB_ &J in~ &RB_ $_|VBG &[obj_N:2]
-       &FAIL to &RB_ @V
C       &[subj_N:1] &AUX_ &RB_ @V~ &[P:0]_ &TO_ &RB_ &V_ &MISC7
C       $to~ &[N:1] to &RB_ &V_ $_ &RB_ &[obj_N:2]
C       $to~ &[N:1] to~ &RB_ &V_ @V~ &RB_ &[N:0] &, &TO_ &RB_ &V_ $_ &RB_
        &[obj_N:2]
-       @EX &AUX [N:0] to &RB_ @V
C       &C1_ [1] [1] -- *** -- to~ &RB_ &V_ $_ &RB_ &[obj_N:2]
C       &C1_ [1] [1] to~ &RB_ &V_ &MISC8

C       &C1_ [1] &MEAN , *** , [2] &WILL [2] &._
C       &C1_ [1] &MEAN [2] &WILL [2] &._

C       &[subj_N:1] touch* off [2] &._
```

F. Subpatterns Used in the Linguistic Patterns

This section lists the set of subpatterns referred to by each subpattern label. The symbol "#" indicates the subpattern label. Below each subpattern label are listed the set of subpatterns for that subpattern label.

The underscore "_" indicates an empty subpattern.

F1. Subpatterns to be used with the patterns in Sections E1 and E2

```
#[1],
[1] , and [1] ,
[1] ,

#[2],
[2] , and [2] ,
[2] ,

#[1](AND_THIS)
[1] &AND &THIS/IT
[1] &AND &THIS [1]
[1]

#[2](AND_THIS)
[2] &AND &THIS/IT
[2] &AND &THIS [2]
[2]

#[N:0]
[N:0]
[M:0]

#[N:1]
[N:1] [P:1] [P:1]
[M:1] [P:1] [P:1]
[N:1] [P:1]
[M:1] [P:1]
[N:1]
[M:1]

#[N:2]
[N:2] [P:2] [P:2]
[M:2] [P:2] [P:2]
[N:2] [P:2]
[M:2] [P:2]
[N:2]
[M:2]

#[conj_N:1]
[1] and [1]

#[conj_N:2]
[2] and [2]

#[conj_N:0],
*** and [N:0] ,
*** and [N:0] [P:0] ,

#[P:0]
[P:0]
@IN &ADJ_ @N @N @N
@IN &ADJ_ @N @N
@IN &ADJ_ @N

#[P:0],
&[P:0] ,
&[P:0] [P:0] ,
&[P:0] [P:0] [P:0] ,
&[P:0] [G:0] ,
&[P:0] [T:0] ,
&[P:0] [P:0] [G:0] ,
&[P:0] [P:0] [T:0] ,

#[N:1],[2]
&[1], &[not_cc:2] [2] &._
[1] [C:2] &.
[1] [C:2] [2] &._
&[N:1] [2] &._

#[not_cc:1]
[N:1]
[M:1]
[cd:1]
[d:1]
[e:1]
[f:1]
[i:1]
[m:1]
[p:1]
[r:1]
[s:1]
[t:1]
[v:1]
[w:1]
" [1]

#[not_cc:2]
[N:2]
[M:2]
[cd:2]
[d:2]
[e:2]
[f:2]
[i:2]
[m:2]
[p:2]
[r:2]
[s:2]
[t:2]
[v:2]
[w:2]
" [2]

#[not_cc:1].
&[not_cc:1] &.
&[not_cc:1] [1] &._

#[not_cc:2].
&[not_cc:2] &.
&[not_cc:2] [2] &._

#[C:1],[2]
&[1], [2] &._
[C:1] [2] &._
[1] [C:2] &.
[1] [C:2] [2] &._

#[C:1].
[C:1] &.
[C:1] [1] &._

#[C:2].
[C:2] &.
[C:2] [2] &._

#[G:1]
```

```
[vbg:1] [1]
[vbg:1]

#[G:2]
[vbg:2] [2]
[vbg:2]

#[G:1].
[vbg:1] &.
[vbg:1] [1] &._

#[G:2].
[vbg:2] &.
[vbg:2] [2] &._

#[C:2]
[C:2]
[2] [v:2] [2]
[2] [v:2]

#[VP:1]
[v:1] [1]
[v:1]

#[VP:2]
[v:2] [2]
[v:2]

#[VP:1].
[v:1] &.
[v:1] [1] &._

#[VP:2].
[v:2] &.
[v:2] [2] &._


#"_
"
_

#,_
, and
,
and
--
_

#,/.
.
,
--

#.
;
:
.
?
, &"_ according~ to~
, &"_ but~
, &"_ and~ [C:0]
, &"_ while~
" $that~
" [M:0] $that~
" [N:0] $that~
, $that~
, [M:0] $that~ &,/.
, [N:0] $that~ &,/.

#._
;
:
, &"_ according~ to~
, &"_ but~
, &"_ and~ [C:0]
, &"_ while~
" $that~
" [M:0] $that~
" [N:0] $that~
, $that~
, [M:0] $that~ &,/.
, [N:0] $that~ &,/.
.
?
_

#A
a
an

#ACHIEVE
achieve
achieved
achieves
achieving

#ADJ
, *** , &DT_ &J
, *** , &DT
&DT_ &J
&DT

#ADJ_
, *** , &DT_ &J
, *** , &DT
&DT_ &J
&DT
_

#ADV
, *** , &RB
&RB

#ADV_
, *** , &RB
&RB
_

#ADVANTAGE
advantage
advantages

#ADVENT
advent
arrival
use
implementation

#AGREE
agree
agreed

#AND
, and
and
; and
;
on one hand , and
on the one hand , and

#AND/OR
and
or

#ANOTHER
another
further
other

#AS/TO_BE_
as
to be
_

#AS/TO_BE/TO_
as
to be
to
_

#AS_AFTER
well
```

```
though
part
in
with
to
if
of
from
for
against
compared
measured
it is
it were
yet
&A
a way
&AUX_ expected
&BE
&HAVE
&INFL
&VBD
[N:0] &BE @vbg
[N:0] $that
many
much
few
$that
president
manager
general
vice*
secretary
head
chairman
chairperson

#AS_BEFORE
just
so
so @J
referred to
rank*
come
came
coming
comes

#ATTRIBUTE
attribute
attributed
attributing
attributes
ascribe
ascribes
ascribed

#AUX
to be
&BE * &CONSIDERED
&AS/TO_BE_
&BE said to be
&INFL * &BECOME
&BE * &BECOME
&INFL * &HAVE * &BE
&INFL * &BE
&HAVE * &BE
&BE
&BECOME
&HAVE to do with
as
--

#AUX_
to be
to
&BE * &CONSIDERED
&AS/TO_BE/TO_
&BE said to be
&BE said to
&INFL * &BECOME
&BE * &BECOME
&INFL * &HAVE * &BE
&INFL * &BE
&INFL * &HAVE
&HAVE * &BE
&BE
&INFL
&BECOME
&HAVE
as
--
_

#BE
is
's|V
are
was
were
am
be
being
been

#BECAUSE/THAT
because
that

#BECOME
becomes
become
became
becoming

#BENEFIT
benefit
benefits
benefited
benefiting

#BLAME
blames
blame
blamed
blaming

#BY_[N:0]_
by [M:0]
by [N:0]
_

#C
@CC~
. &"_
; &"_
: &"_
because~ &,_ &"_
that|IN~ &,_ &"_
after~ all~ , &"_
according~ to~ &[N:0] ,
&"_
also~ , &"_
in~ addition~ , &"_
however~ , &"_
recently~ , &"_
, &"_ &IN_ &TIME , &"_
, &"_ @CC~ &TIME , &"_
, &"_ but~ &"_
, &"_ while~ &"_
[C:0] &"_ , and~ &"_
$that~ &MISC1
&WH &"_
. &"_ &CC_ so~ &,_ &"_
. &"_ &CC_ if~ so~ &,_
&"_
. &"_ &CC_ &RB , &"_
. &"_ &CC_ &RB &"_
. &"_ &CC_ [P:0] , &"_
. &"_ &CC_ [P:0] [P:0] ,
&"_
. &"_ &CC_ @IN &TIME
```

```
&"_
. &"_ @CC~ &TIME , &"_
. &"_ @CC~ &,_ &"_
. &"_ &CC_ [G:0] , &"_
. &"_ &CC_ [G:0] [P:0] ,
&"_

#C_
;
:
because~
that|IN~
after~ all~ , &"_
according~ to~ &[N:0] ,
also~ ,
in~ addition~ ,
however~ ,
recently~ ,
, &"_ &IN_ &TIME ,
, &"_ @CC~ &TIME ,
, &"_ but~
, &"_ while~
[C:0] , and~
$that~ &MISC2
&WH
. &"_ &CC_ so~
. &"_ &CC_ if~ so~
. &"_ &CC_ &RB ,
. &"_ &CC_ &RB
. &"_ &CC_ [P:0] ,
. &"_ &CC_ [P:0] [P:0] ,
&"_
. &"_ &CC_ @IN &TIME
. &"_ @CC~ &TIME ,
. &"_ @CC~
. &"_ &CC_ [G:0] ,
. &"_ &CC_ [G:0] [P:0] ,
_

#C1
; &"_
: &"_
because~ &"_
that|IN~ &"_
[1] &AUX &RB_ $that~
&BY_[N:0]_
after~ all~ , &"_
according~ to~ &[N:0] ,
&"_
also~ , &"_
in~ addition~ , &"_
however~ , &"_
recently~ , &"_
[1] , &"_ &IN_ &TIME ,
&"_
, &"_ @CC~ &TIME , &"_
, &"_ but~ &"_
, &"_ while~ &"_
[C:0] &"_ , and~ &"_
[1] , *** $that~ , &"_
[1] , $that~ *** , &"_
$that~ &MISC1
&WH &"_
[1] @V *** &AND/OR
&AUX_ &RB_ [v:1]
. &"_ &CC_ so~ &,_ &"_
. &"_ &CC_ if~ so~ &,_
&"_
. &"_ &CC_ &RB , &"_
. &"_ &CC_ &RB &"_
. &"_ &CC_ [P:0] , &"_
. &"_ &CC_ [P:0] [P:0] ,
&"_
. &"_ &CC_ @IN &TIME
&"_
. &"_ @CC~ &TIME , &"_
. &"_ @CC~ &,_ &"_
. &"_ &CC_ [G:0] , &"_
. &"_ &CC_ [G:0] [P:0] ,
&"_

#C1_
;
:
because~
that|IN~
[1] &AUX &RB_ $that~
&BY_[N:0]_
after~ all~ , &"_
according~ to~ &[N:0] ,
also~ ,
in~ addition~ ,
however~ ,
recently~ ,
[1] , &IN_ &TIME ,
, &"_  @CC~ &TIME ,
, &"_ but~
, &"_ while~
[C:0] , and~
[1] , *** $that~ ,
[1] , $that~ *** ,
$that~ &MISC2
&WH
[1] @V *** &AND/OR
&AUX_ &RB_ [v:1]
. &"_ &CC_ so~
. &"_ &CC_ if~ so~
. &"_ &CC_ &RB ,
. &"_ &CC_ &RB
. &"_ &CC_ [P:0] ,
. &"_ &CC_ [P:0] [P:0] ,
&"_
. &"_ &CC_ @IN &TIME
. &"_ @CC~ &TIME ,
. &"_ @CC~
. &"_ &CC_ [G:0] ,
. &"_ &CC_ [G:0] [P:0] ,
_

#C2
; &"_
: &"_
because~ &"_
that|IN~ &"_
[2] &AUX &RB_ $that~
&BY_[N:0]_
after~ all~ , &"_
according~ to~ &[N:0] ,
&"_
also~ , &"_
in~ addition~ , &"_
however~ , &"_
recently~ , &"_
[2] , &"_ &IN_ &TIME ,
&"_
, &"_ @CC~ &TIME , &"_
, &"_ but~ &"_
, &"_ while~ &"_
[C:0] &"_ , and~ &"_
[2] , *** $that~ , &"_
[2] , $that~ *** , &"_
$that~ &MISC1
&WH &"_
[2] @V *** &AND/OR
&AUX_ &RB_ [v:2]
. &"_ &CC_ so~ &,_ &"_
. &"_ &CC_ if~ so~ &,_
&"_
. &"_ &CC_ &RB , &"_
. &"_ &CC_ &RB &"_
. &"_ &CC_ [P:0] , &"_
. &"_ &CC_ [P:0] [P:0] ,
&"_
. &"_ &CC_ @IN &TIME
&"_
. &"_ @CC~ &TIME , &"_
. &"_ @CC~ &,_ &"_
. &"_ &CC_ [G:0] , &"_
. &"_ &CC_ [G:0] [P:0] ,
&"_

#C2_
```

```
;
:
because~
that|IN~
[2] &AUX &RB_ $that~
&BY_[N:0]_
after~ all~ , &"_
according~ to~ &[N:0] ,
also~ ,
in~ addition~ ,
however~ ,
recently~ ,
[2] , &IN_ &TIME ,
, &"_ @CC~ &TIME ,
, &"_ but~
, &"_ while~
[C:0] , and~
[2] , *** $that~ ,
[2] , $that~ *** ,
$that~ &MISC2
&WH
[2] @V *** &AND/OR
&AUX_ &RB_ [v:2]
. &"_ &CC_ so~
. &"_ &CC_ if~ so~
. &"_ &CC_ &RB ,
. &"_ &CC_ &RB
. &"_ &CC_ [P:0] ,
. &"_ &CC_ [P:0] [P:0] ,
&"_
. &"_ &CC_ @IN &TIME
. &"_ @CC~ &TIME ,
. &"_ @CC~
. &"_ &CC_ [G:0] ,
. &"_ &CC_ [G:0] [P:0] ,
_


#C/,
,
@CC
, @CC
&C

#CAUSE
cause
causes

#CC_
@CC
_

#CONSIDERED
considered
thought
cited

#CONTRIBUTE
contribute
contributed
contributes
contributing

#CRITICAL
critical
vital
essential
important
imperative
crucial

#DELIVER
deliver
delivers
delivered
delivering
&PROVIDE

#DISADVANTAGE
disadvantages
disadvantage

#DT
@DT
@CD
its

#DT_
@DT
@CD
its
_

#DUE_TO
due to
owing to
depend on
dependent on
depends on
depending on
depended on
contingent on
attributed to
attributable to

#EFFECT
effect
effects
impact
influence

#EFFECT/RESULT
effect
effects
result
results
consequence
consequences
impact
influence

#EMERGE
emerge
emerged
emerging
emerges

#ENABLE
enable
enables
enabled
enabling
permit
permits
permitted
permitting
allow
allows
allowed
allowing

#FACTOR
factor
factors
determinant
determinants

#FOR
for
FOR

#FOR/WHY
for
why

#GET
get
gets
got

#GROUND
ground
```

grounds

#HAVE
have
has
had
having
've
'd

#HOLD
hold
holds
holding
held

#IF
if , *** ,
if

#IN_
@IN
_

#INCLUDE
include
includes
included
including
&BE

#INFL
may
might
should
must
will
shall
might
would
can
could
wo
do
did

#INITIATOR
initiator
initiators

#J
&RB_ @J &,_ &RB_ @J
&,_ &RB_ @J
&RB_ @J &,_ &RB_ @J
&RB_ @J

#J_
&RB_ @J &,_ &RB_ @J
&,_ &RB_ @J
&RB_ @J &,_ &RB_ @J
&RB_ @J
_

#KNOW
know
knew
knowing
knows
known

#LOOKING
looking
wanting
intending

#MAKE
make
makes
made
making

#MANY
many
several
few
number of
some
diverse
different
various
numerous
@CD

#MISC1
, &"_
[M:0] , &"_
[N:0] , &"_
&"_

#MISC2
[M:0] ,
[N:0] ,
_

#MISC3
that
&,_ &"_

#MUST
must
have to
had to
has to
need to
needs to
needed to
should
better|RB

#NECESSARY
necessary
mandatory
obligatory

#NO
no
little
hardly any
minimal

#NOT
not
n't
unlikely to

#NOUN_FOR
rule
rules
motive
motives
proposal
potential
case
plan
plans
incentive
incentives
solution
solutions
prospect*
deadline

#NOUN_TO
tendency
plan
plans
strategy
opportunity
reason
likely
enough
means

```
delivery
deliveries
ability
attempt
measure
measures
way

#OBSTACLE
obstacle
obstacles
barrier
barriers
impediment
impediments

#OF/FOR
of
for

#OF/FROM
of
from

#OF/FROM/TO
of
from
to

#OF/IN
of
in

#OF/ON
of
on

#PERSON
&"_ @NP
&"_ @PP
&"_ president
&"_ presidents
&"_ man
&"_ men
&"_ woman
&"_ women
&"_ person
&"_ persons
&"_ businessman
&"_ businessmen
&"_ executive
&"_ executives
&"_ manager
&"_ managers
&"_ buyer
&"_ buyers
&"_ seller
&"_ sellers
&"_ client
&"_ clients

#PLAY
play
played
plays
playing

#PREDICTOR
predictor
predictors

#PROMISE
promise
promises
promised
promising

#PROVIDE
provide
provides
provided
providing

#RB
@RB &,_ @RB &,_ @RB
@RB &,_ @RB
@RB

#RB_
@RB &,_ @RB &,_ @RB
@RB &,_ @RB
@RB
_

#REASON
reason
reasons
explanation
explanations
&GROUND

#REPORT
report
reports
reported
reporting

#REQUIRE
requires
required
require
requiring

#RESPOND
respond
responds
responded

#RESULT
result
results
consequence
consequences

#S
. &"_
; &"_
: &"_
. after~ all~ , &"_
. according~ to~ &[N:0] ,
&"_
. also~ , &"_
. in~ addition , &"_
. however , &"_
. recently , &"_
. &[N:0] $that &MISC3
. $that &[N:0] , &"_
. &"_ so &,_ &"_
. &"_ if so &,_ &"_
. &"_ &RB , &"_
. &"_ [P:0] , &"_
. &"_ [P:0] [P:0] , &"_
. &"_ @IN &TIME &"_
. &"_ @CC &TIME , &"_
. &"_ @CC &,_ &"_

#S_
after~ all~ , &"_
according~ to~ &[N:0] ,
&"_
also~ , &"_
in~ addition~ , &"_
however~ , &"_
recently~ , &"_
&[N:0] $that~ &MISC3
$that~ &[N:0] , &"_
&"_ so~ &,_ &"_
&"_ if~ so~ &,_ &"_
&"_ &RB , &"_
&"_ [P:0] , &"_
&"_ [P:0] [P:0] , &"_
&"_ @IN &TIME &"_
```

&"_ @CC~ &TIME , &"_
&"_ @CC~ &,_ &"_
"
_

#SAME/DIFFERENT
same
similar
different

#SAY
say
says
said

#SERVE
serve
serves
served
serving

#STAND
stand
stands
stood
standing

#TAKE
take
took
takes
taken
taking

#THERE
there
here
the following

#THEREFORE
therefore
thus
consequently
hence
as a result
accordingly
as a consequence

#THINK
think
thinks
thought

#THIS
this
these
that
those

#THIS/HERE
this
these
that
those
here

#THIS/IT
this
these
that
those
it

#THIS,[2]
&THIS , [2]
&THIS [C:2]
&THIS *** , [2]
&THIS *** [C:2]
&THIS [2]

#THIS[1],[2]
&THIS , [2]
&THIS [C:2]
&THIS [1] , [2]
&THIS [1] [C:2]
&THIS [2]

#THROUGH/TO
through *** and ***
through
through *** but ***
through
through
&TO

#TIME
&DT_ beginning of &ADJ_ &TIME2
&DT_ end of &ADJ_ &TIME2
&DT_ middle of &ADJ_ &TIME2
&DT_ rest of &ADJ_ &TIME2
all of &ADJ_ &TIME2
&ADJ_ half of &ADJ_ &TIME2
&ADJ_ quarter of &ADJ_ &TIME2
most of &ADJ_ &TIME2
much of &ADJ_ &TIME2
&ADJ_ &TIME2

#TIME2
year*
quarter*
half
month*
period
week*
day*
hour*
minute*
second*
today
tomorrow
yesterday
tonight
noon
midnight
morning*
afternoon*
evening*
night*
SPRING
SUMMER*
FALL
WINTER*
spring
summer*
fall
winter*
JAN.*
JANUARY
FEB.*
FEBRUARY
MAR.*
MARCH
APR.*
APRIL
MAY
JUNE
JUL.*
JULY
AUG.*
AUGUST
SEP.*
SEPT.*
SEPTEMBER
OCT.*
OCTOBER
NOV.*

NOVEMBER
DEC.*
DECEMBER
MON.*
MONDAY
TUE.*
TUESDAY
WED.*
WEDNESDAY
THU.*
THUR.*
THURSDAY
FRI.*
FRIDAY
SAT.*
SATURDAY
SUN.*
SUNDAY
0*
1*
2*
3*
4*
5*
6*
7*
8*
9*
mid*

#TO
to *** and &RB_ to
to *** but &RB_ to
to

#TO/FOR
to
for

#TREND
trend
trends

#VBD
@VBD
@VBN

#VBD/VBG
@VBD
@VBN
@VBG

#VERB_BY
respond*
conclude*
beguile*
repay*

#VERB_FOR
call
calls
called
keep
keeps
kept
use
used
uses

#WANT
want
wants
wanted
wanting
mean
means
meant
meaning
&BE

#WAY
way
ways

#WH
what
where
how
why
when
whether

#WHICH
which
that
who

#WILL
will
would
may
might
shall
can
could
wo
'll
'd
plan* to
stand* to
$BE likely

F1. Subpatterns to be used with the patterns in Section E3

#past_v+against
$+against|VBD
$+against|VBN

#past_v+among
$+among|VBD
$+among|VBN

#past_v+around
$+around|VBD
$+around|VBN

#past_v+as
$+as|VBD
$+as|VBN

#past_v+at
$+at|VBD
$+at|VBN

#past_v+back_on
$+back_on|VBD
$+back_on|VBN

#past_v+back_to
$+back_to|VBD
$+back_to|VBN

#past_v+between
$+between|VBD
$+between|VBN

#past_v+for
$+for|VBD
$+for|VBN

#past_v+from
$+from|VBD
$+from|VBN

#past_v+from_v-ing
$+from|VBD
$+from|VBN

#past_v+in_favor_of
$+in_favor_of|VBD
$+in_favor_of|VBN

#past_v+in_for
$+in_for|VBD
$+in_for|VBN

#past_v+in_on
$+in_on|VBD
$+in_on|VBN

#past_v+in
$+in|VBD
$+in|VBN

#past_v+into
$+into|VBD
$+into|VBN

#past_v+of
$+of|VBD
$+of|VBN

#past_v+off_to
$+off_to|VBD
$+off_to|VBN

#past_v+off_with
$+off_with|VBD
$+off_with|VBN

#past_vdown
$down|VBD
$down|VBN

#past_vdown_on
$down_on|VBD
$down_on|VBN

#past_v+on
$+on|VBD
$+on|VBN

#past_v+onto
$+onto|VBD
$+onto|VBN

#past_v+out_in
$+out_in|VBD
$+out_in|VBN

#past_v+out_of
$+out_of|VBD
$+out_of|VBN

#past_v+out
$+out|VBD
$+out|VBN

#past_v+over_to
$+over_to|VBD
$+over_to|VBN

#past_v+over
$+over|VBD
$+over|VBN

#past_v+round
$+round|VBD
$+round|VBN

#past_v+through
$+through|VBD
$+through|VBN

#past_v+throughout
$+throughout|VBD
$+throughout|VBN

#past_v+to-v
$+to-v|VBD
$+to-v|VBN

#past_v+to
$+to|VBD
$+to|VBN

#past_v+towards
$+towards|VBD
$+towards|VBN

#past_v+upon
$+upon|VBD
$+upon|VBN

#past_v+with
$+with|VBD
$+with|VBN

#past_vabout
$about|VBD
$about|VBN

#past_vapart
$apart|VBD

```
$apart|VBN

#past_vaway
$away|VBD
$away|VBN

#past_vback
$back|VBD
$back|VBN

#past_vdown
$down|VBD
$down|VBN

#past_vforth
$forth|VBD
$forth|VBN

#past_vin
$in|VBD
$in|VBN

#past_voff
$off|VBD
$off|VBN

#past_von
$on|VBD
$on|VBN

#past_vopen
$open|VBD
$open|VBN

#past_vout
$out|VBD
$out|VBN

#past_vover
$over|VBD
$over|VBN

#past_vround
$round|VBD
$round|VBN

#past_vthrough
$through|VBD
$through|VBN

#past_vtogether
$together|VBD
$together|VBN
```

```
#past_vup
$up|VBD
$up|VBN

#past_vphr_away_with
$phr_away_with|VBD
$phr_away_with|VBN

#past_vphr_from
$phr_from|VBD
$phr_from|VBN

#past_vphr_in
$phr_in|VBD
$phr_in|VBN

#past_vphr_out+to
$phr_out+to|VBD
$phr_out+to|VBN

#past_v+n
$+n|VBD
$+n|VBN

#past_v+prep
$+prep|VBD
$+prep|VBN

#past_v+v-ing
$+v-ing|VBD
$+v-ing|VBN

#past_v+v-ed
$+v-ed|VBD
$+v-ed|VBN

#past_v+v
$+v|VBD
$+v|VBN

#past_vparticle
$particle|VBD
$particle|VBN

#past_v_
$_|VBD
$_|VBN


#[1],
[1] , and [1] ,
[1] ,

#[2],
```

```
[2] , and [2] ,
[2] ,

#[1](AND_THIS)
[1] &AND &THIS/IT
[1] &AND &THIS [1]
[1]

#[2](AND_THIS)
[2] &AND &THIS/IT
[2] &AND &THIS [2]
[2]

#[N:0]
[M:0] [P:0] [P:0]
[N:0] [P:0] [P:0]
[M:0] [P:0]
[N:0] [P:0]
[M:0] [G:0]
[N:0] [G:0]
[M:0]
[N:0]

#[N:0]_
_
[M:0]
[N:0]
[M:0] [P:0] [P:0]
[N:0] [P:0] [P:0]
[M:0] [P:0]
[N:0] [P:0]

#[N:1]
[M:1] [P:1] [P:1]
[N:1] [P:1] [P:1]
[M:1] [P:1]
[N:1] [P:1]
[M:1]
[N:1]

#[N:2]
[M:2] [P:2] [P:2]
[N:2] [P:2] [P:2]
[M:2] [P:2]
[N:2] [P:2]
[M:2]
[N:2]


#[subj_N:1]
&C_ [1] , and that
&[N:1] &SAY &THAT_
@PP~
[G:1] &MISC4_ &"_
```

```
&C_ &[N:1] &MISC4_
&"_

#[subj_N:1]ver2
&C_ [1] , and that
&[N:1] &SAY &THAT_
@PP~
[G:1] &MISC4_ &"_
&[N:1] &MISC4b &"_
&C_ &[N:1] &,***,_

#[subj_N:1]ver1
&C_ [1] , and that
&[N:1] &SAY &THAT_
@PP~
&C_ [G:1] &MISC4_ &"_
&C_ &[N:1] &MISC4_
&"_


#[subj_N:2]
&C_ [2] , and that
&[N:2] &SAY &THAT_
@PP~
&C_ &[N:2] &MISC4_
&"_

#[subj_N:2]ver2
&C_ [2] , and that
&[N:2] &SAY &THAT_
@PP~
&[N:2] &MISC4b &"_
&C_ &[N:2] &,***,_

#[subj_N:2]ver1
&C_ [2] , and that
&[N:2] &SAY &THAT_
@PP~
&C_ [G:2] &MISC4_ &"_
&C_ &[N:2] &MISC4_
&"_

#[obj_N:1]
&[N:1]
@CC~ &AUX_ &RB_
@V~ &[N:1]

#[obj_N:2]
&[N:2]
@CC~ &AUX_ &RB_
@V~ &[N:2]

#[conj_N:1]
[1] and [1]

#[conj_N:2]
[2] and [2]

#[conj_N:0],
*** and [N:0] ,
*** and [N:0] [P:0] ,

#[P:0]
[P:0]
@IN &ADJ_ @N @N @N
@IN &ADJ_ @N @N
@IN &ADJ_ @N

#[P:0]_
_
[P:0]
[P:0] [P:0]

#[P:1]
[P:1]

#[P:2]
[P:2]

#[P:0],
&[P:0] ,
&[P:0] [P:0] ,
&[P:0] [P:0] [P:0] ,
&[P:0] [G:0] ,
&[P:0] [T:0] ,
&[P:0] [P:0] [G:0] ,
&[P:0] [P:0] [T:0] ,

#[N:1],[2]
&[1], &"_ [not_cc:2] [2]
&._
[1] [C:2] &.
[1] [C:2] [2] &._
&[N:1] [2] &._


#[C:1],[2]
&[1], [2] &._
[C:1] [2] &._
[1] [C:2] &.
[1] [C:2] [2] &._

#[C:1].
[C:1] &.
[C:1] [1] &._

#[C:2].
[C:2] &.
[C:2] [2] &._

#[G:1]
[vbg:1] [1]
[vbg:1]

#[G:2]
[vbg:2] [2]
[vbg:2]

#[G:1].
[vbg:1] &.
[vbg:1] [1] &._

#[G:2].
[vbg:2] &.
[vbg:2] [2] &._

#[C:2]
[C:2]
[2] [v:2] [2]
[2] [v:2]

#[VP:1]
[v:1] [1]
[v:1]

#[VP:2]
[v:2] [2]
[v:2]

#[VP:1].
[v:1] &.
[v:1] [1] &._

#[VP:2].
[v:2] &.
[v:2] [2] &._

#[VPpast:1]
[vbd:1] [1]
[vbn:1] [1]
[vbd:1]
[vbn:1]

#[VPpast:2]
[vbd:2] [2]
[vbn:2] [2]
[vbd:2]
[vbn:2]


#"_
"
```

```
_

#,
, and~
,
and~

#,_
, and~
,
and~
--
(
_

#,***,_
, *** ,
_

#,/.
.
,
--

#.
;
:
.
?
, &"_ according to
, &"_ but
, &"_ and [C:0]
, &"_ while
" $that
" [M:0] $that
" [N:0] $that
, $that
, [M:0] $that &,/.
, [N:0] $that &,/.

#._
;
:
, &"_ according to
, &"_ but
, &"_ and [C:0]
, &"_ while
" $that
" [M:0] $that
" [N:0] $that
, $that
, [M:0] $that &,/.
, [N:0] $that &,/.
.
?
_

#A
a
an

#ADJ
, *** , &DT_ &J
, *** , &DT
&DT_ &J
&DT

#ADJ_
, *** , &DT_ &J
, *** , &DT
&DT_ &J
&DT
_

#ADMIT
admit
admits
admitted
admitting

#ADV
, *** , &RB
&RB

#ADV_
, *** , &RB
&RB
_

#AFFECT
affect
affects
affected
affecting

#AND
, and
and
; and
;
on one hand , and
on the one hand , and

#AND/OR
and
or

#AS/TO_BE_
as
to be
_

#AS/TO_BE/TO_
as
to be
to
_

#ASSURE
assure
assures
assured
assuring

#AUX
to be
&BE * &CONSIDERED
&AS/TO_BE_
&BE said to be
&INFL * &BECOME
&BE * &BECOME
&INFL * &HAVE * &BE
&INFL * &BE
&HAVE * &BE
&BE
&BECOME
&HAVE to do with
as
for
by
--
,

#AUX_
to be
&BE * &CONSIDERED
&AS/TO_BE/TO_
&BE said to be
&BE said to
&INFL * &BECOME
&BE * &BECOME
&INFL * &HAVE * &BE
&INFL * &BE
&INFL * &HAVE
&HAVE * &BE
&BE
&INFL
&BECOME
&HAVE
as
```

```
--
,
_

#BAR
bar
barred
barring
bars

#BE
is
's|V
are
was
were
am
be
being
been

#BE_
be
become
_

#BECOME
becomes
become
became
becoming

#BY_[N:0]_
by [M:0]
by [N:0]
_

#BRING
bring
brings
brought
bringing

#BUGGER
bugger
buggers
buggered

#C
@CC
. &"_
; &"_
: &"_
because~ &,_ &"_
that|IN &"_
after all , &"_
according to &[N:0] , &"_
also , &"_
in addition , &"_
however , &"_
recently , &"_
, &"_ &IN_ &TIME , &"_
, &"_ @CC &TIME , &"_
, &"_ but &"_
, &"_ and &"_
, &"_ while &"_
[C:0] &"_ , and &"_
$that &MISC1
&WH &"_
. &"_ so &,_ &"_
. &"_ if so &,_ &"_
. &"_ &RB , &"_
. &"_ &RB &"_
. &"_ [P:0] , &"_
. &"_ [P:0] [P:0] , &"_
. &"_ @IN &TIME &"_
. &"_ @CC &TIME , &"_
. &"_ @CC &,_ &"_
. &"_ [G:0] , &"_
. &"_ [G:0] [P:0] , &"_

#C_
;
:
because~
that|IN
after all , &"_
according to &[N:0] ,
also ,
in addition ,
however ,
recently ,
, &"_ &IN_ &TIME ,
, &"_ @CC &TIME ,
, &"_ but
, &"_ and
, &"_ while
[C:0] , and
&WH
. &"_ so
. &"_ if so
. &"_ &RB ,
. &"_ &RB
. &"_ [P:0] ,
. &"_ [P:0] [P:0] , &"_
. &"_ @IN &TIME
. &"_ @CC &TIME ,
. &"_ &"_ @CC
. &"_ [G:0] ,
. &"_ [G:0] [P:0] ,
_

#C1
; &"_
: &"_
[1] &AUX $that
&BY_[N:0]_
because~ &"_
that|IN &"_
after all , &"_
according to &[N:0] , &"_
also , &"_
in addition , &"_
however , &"_
recently , &"_
[1] , &"_ &IN_ &TIME ,
&"_
, &"_ @CC &TIME , &"_
, &"_ but &"_
, &"_ while &"_
[C:0] &"_ , and &"_
[1] , *** $that , &"_
[1] , $that *** , &"_
$that &MISC1
&WH &"_
[1] @V *** &AND/OR
&AUX_ &RB_ [v:1]
. &"_ so &,_ &"_
. &"_ if so &,_ &"_
. &"_ &RB , &"_
. &"_ &RB &"_
. &"_ [P:0] , &"_
. &"_ [P:0] [P:0] , &"_
. &"_ @IN &TIME &"_
. &"_ @CC &TIME , &"_
. &"_ @CC &,_ &"_
. &"_ [G:0] , &"_
. &"_ [G:0] [P:0] , &"_

#C1_
;
:
[1] &AUX $that
&BY_[N:0]_
because~
that|IN
after all , &"_
according to &[N:0] ,
also ,
in addition ,
however ,
recently ,
[1] , &IN_ &TIME ,
, &"_  @CC &TIME ,
, &"_ but
, &"_ while
[C:0] , and
[1] , *** $that ,
[1] , $that *** ,
$that &MISC2
&WH
[1] @V *** &AND/OR
&AUX_ &RB_ [v:1]
. &"_ so
. &"_ if so
. &"_ &RB ,
. &"_ &RB
. &"_ [P:0] ,
. &"_ [P:0] [P:0] , &"_
. &"_ @IN &TIME
. &"_ @CC &TIME ,
. &"_ @CC
. &"_ [G:0] ,
. &"_ [G:0] [P:0] ,
_

#C2
; &"_
: &"_
[2] &AUX $that
&BY_[N:0]_
because~ &"_
that|IN &"_
after all , &"_
according to &[N:0] , &"_
also , &"_
in addition , &"_
however , &"_
recently , &"_
[2] , &"_ &IN_ &TIME ,
&"_
, &"_ @CC &TIME , &"_
, &"_ but &"_
, &"_ while &"_
[C:0] &"_ , and &"_
[2] , *** $that , &"_
[2] , $that *** , &"_
$that &MISC1
&WH &"_
[2] @V *** &AND/OR
&AUX_ &RB_ [v:2]
. &"_ so &,_ &"_
. &"_ if so &,_ &"_
. &"_ &RB , &"_
. &"_ &RB &"_
. &"_ [P:0] , &"_
```

```
. &"_ [P:0] [P:0] , &"_
. &"_ @IN &TIME &"_
. &"_ @CC &TIME , &"_
. &"_ @CC &,_ &"_
. &"_ [G:0] , &"_
. &"_ [G:0] [P:0] , &"_

#C2_
;
:
[2] &AUX $that
&BY_[N:0]_
because~
that|IN
after all , &"_
according to &[N:0] ,
also ,
in addition ,
however ,
recently ,
[2] , &IN_ &TIME ,
, &"_ @CC &TIME ,
, &"_ but
, &"_ while
[C:0] , and
[2] , *** $that ,
[2] , $that *** ,
$that &MISC2
&WH
[2] @V *** &AND/OR
&AUX_ &RB_ [v:2]
. &"_ so
. &"_ if so
. &"_ &RB ,
. &"_ &RB
. &"_ [P:0] ,
. &"_ [P:0] [P:0] , &"_
. &"_ @IN &TIME
. &"_ @CC &TIME ,
. &"_ @CC
. &"_ [G:0] ,
. &"_ [G:0] [P:0] ,
_


#C/,
,
@CC
, @CC
&C

#CHANGE_HAND
change hands
changed hands
```

```
changing hands
changes hands

#CONCLUDE
conclude
concluded
concludes
concluding

#CONSIDERED
considered
thought
cited

#DT
@DT
@CD
its

#DT_
@DT
@CD
its
_


#EXPRESS
express
expressed
expressing

#FAIL
fail
fails
failed
failing

#FILE
file
filed
files
filing

#GIVE
give
gives
gave
given
giving

#HAVE
have
has
had
```

```
having
've
'd

#HOLD
hold
holds
held
holding

#IN_
@IN
_

#INCLUDE
include
includes
included
including

#INFL
may
might
should
must
will
shall
might
would
can
could
wo
do
did

#INTERSPERSE
intersperse
intersperses
interspersed
interspersing

#J
&RB_ @J &,_ &RB_ @J
&,_ &RB_ @J
&RB_ @J &,_ &RB_ @J
&RB_ @J

#J_
&RB_ @J &,_ &RB_ @J
&,_ &RB_ @J
&RB_ @J &,_ &RB_ @J
&RB_ @J
_

#LEAD
lead
leads
led
leading

#LET
let
lets
letting

#MAINTAIN
maintain
maintains
maintained
maintaining

#MAKE
make
makes
made
making

#MANY
many
several
few
number of
some
diverse
different
various
numerous
@CD

#MEAN
mean
means
meant


#MISC1
, &"_
[M:0] , &"_
[N:0] , &"_
&"_

#MISC2
[M:0] ,
[N:0] ,
_

#MISC3
that
&,_ &"_

#MISC4_
more than anything else
&,_ &WHICH
&WHICH , *** ,
&,_ &WHICH &,_ [N:0]
&SAY &,_
&,_ &WHICH_ * * @V~
&[N:0]_ &,
XXX &,_ &WHICH_ * *
@V~ &[N:0]_ , * * @V~
&[N:0]_ &,
-- *** --
, *** ,
_

#MISC4b
more than anything else
&,_ &WHICH
&WHICH , *** ,
&,_ &WHICH &,_ [N:0]
&SAY &,_
&,_ &WHICH_ * * @V~
&[N:0]_ &,
XXX &,_ &WHICH_ * *
@V~ &[N:0]_ , * * @V~
&[N:0]_ &,
-- *** --

#MISC4_old
more than anything else
&,_ &WHICH
&WHICH , *** ,
&,_ &WHICH &,_ [N:0]
&SAY &,_
&,_ &WHICH_ &AUX_
&RB_ @V~ &[N:0]_ &,
&,_ &WHICH_ &AUX_
&RB_ @V~ &[N:0]_ ,
&AUX_ &RB_ @V~
&[N:0]_ &,
-- *** --
, *** ,
_

#MISC5
[I:2]
@CC &RB_ [j:2] &"_
&NOT_NOUN
&"_ &NOT_NOUN

#MISC7
$_ &RB_ &[obj_N:2]
```

```
@V~ &RB_ &[N:0] &,
&TO_ &RB_ &V_ $_
&RB_ &[obj_N:2]

#MISC8
$_ &RB_ &[obj_N:2]
@V~ &RB_ &[N:0] &,
&TO_ &RB_ &V_ $_
&RB_ &[obj_N:2]

#MISC9
to~ &RB_ &V_ $_
&[obj_N:2]
to~ &RB_ &V_ @V~
&[N:0] &, &TO_ &RB_
&V_ $_ &[obj_N:2]

#MISC10
&[N:2] out~ to~ &[N:2]
out~ &[N:2] to~ &[N:2]

#MISC11
&[N:2] up~
up~ &[N:2]

#MISC12
&[N:2] together~
together~ &[N:2]

#MISC13
&[N:2] through~
through~ &[N:2]

#MISC14
&[N:2] round~
round~ &[N:2]

#MISC15
&[N:2] over~
over~ &[N:2]

#MISC16
&[N:2] out~
out~ &[N:2]

#MISC17
&[N:2] open~
open~ &[N:2]

#MISC18
&[N:2] on~
on~ &[N:2]

#MISC19
&[N:2] off~
off~ &[N:2]

#MISC20
&[N:2] in~
in~ &[N:2]

#MISC21
&[N:2] forth~
forth~ &[N:2]

#MISC22
&[N:2] down~
down~ &[N:2]

#MISC23
&[N:2] back~
back~ &[N:2]

#MISC24
&[N:2] away~
away~ &[N:2]

#MISC25
&[N:2] apart~
apart~ &[N:2]

#MISC26
&[N:2] about~
about~ &[N:2]

#MISC27
that~ [2] &._
[C:2]

#MISC28
[2] down~ on~ &[N:2]
down~ &[N:2] on~ &[N:2]

#MISC31
back~ &[N:2] to~ &[N:2]
&[N:2] back~ to~ &[N:2]

#MISC40
@RP &[N:2]
&[N:2] @RP

#MUST
must
have to
had to
has to

#NO
no
little
hardly any
minimal

#NOT
not
n't
unlikely to

#NOT_J
such
more

#NOT_NOUN
@CD
@DT
@EX
@IN
@MD
@P
@W
&.
&,_ &J &.
&, &"_ [N:0]
&, &"_ [P:0]
&, &"_ @V~
--

#OF_
_
of

#OFFER
offer
offered
offers
offering

#OPEN_FIRE
open fire
opens fire
opened fire

#PERSON
&"_ @NP
&"_ @PP
&"_ president
&"_ presidents
&"_ man
&"_ men
&"_ woman
&"_ women
&"_ person
```

```
&"_ persons
&"_ businessman
&"_ businessmen
&"_ executive
&"_ executives
&"_ manager
&"_ managers
&"_ buyer
&"_ buyers
&"_ seller
&"_ sellers
&"_ client
&"_ clients

#POSE
pose
poses
posed
posing

#POST
post
posts
posting
posted

#QUESTION
question
questions

#RAISE_QUESTION
raise * &QUESTION
raises * &QUESTION
raised * &QUESTION

#RB
@RB~ &,_ @RB~ &,_
@RB~
@RB~ &,_ @RB~
@RB~

#RB_
@RB~ &,_ @RB~ &,_
@RB~
@RB~ &,_ @RB~
@RB~
_

#S
. &"_
; &"_
: &"_
. after all , &"_
. according to &[N:0] , &"_
. also , &"_
. in addition , &"_
. however , &"_
. recently , &"_
. &[N:0] $that &MISC3
. $that &[N:0] , &"_
. &"_ so &,_ &"_
. &"_ if so &,_ &"_
. &"_ &RB , &"_
. &"_ [P:0] , &"_
. &"_ [P:0] [P:0] , &"_
. &"_ @IN &TIME &"_
. &"_ @CC &TIME , &"_
. &"_ @CC &,_ &"_

#S_
after all , &"_
according to &[N:0] , &"_
also , &"_
in addition , &"_
however , &"_
recently , &"_
&[N:0] $that &MISC3
$that &[N:0] , &"_
&"_ so &,_ &"_
&"_ if so &,_ &"_
&"_ &RB , &"_
&"_ [P:0] , &"_
&"_ [P:0] [P:0] , &"_
&"_ @IN &TIME &"_
&"_ @CC &TIME , &"_
&"_ @CC &,_ &"_
"
_

#SAVE_TIME
save time
saves time
saved time
saving time

#SAY
say
says
said

#SEND
send
sends
sent
sending

#SET
set
sets
setting

#TAKE
take
took
taken
takes
taking

#THAT_
that
_

#THIS
this
these
that
those

#THIS/IT
this
these
that
those
it


#TIME
&DT_ beginning of &ADJ_
&TIME2
&DT_ end of &ADJ_
&TIME2
&DT_ middle of &ADJ_
&TIME2
&DT_ rest of &ADJ_
&TIME2
all of &ADJ_ &TIME2
&ADJ_ half of &ADJ_
&TIME2
&ADJ_ quarter of &ADJ_
&TIME2
most of &ADJ_ &TIME2
much of &ADJ_ &TIME2
&ADJ_ &TIME2

#TIME2
year*
quarter*
half
month*
period
week*
day*
```

hour*
minute*
second*
today
tomorrow
yesterday
tonight
noon
midnight
morning*
afternoon*
evening*
night*
SPRING
SUMMER*
FALL
WINTER*
spring
summer*
fall
winter*
JAN.*
JANUARY
FEB.*
FEBRUARY
MAR.*
MARCH
APR.*
APRIL
MAY
JUNE
JUL.*
JULY
AUG.*
AUGUST
SEP.*
SEPT.*
SEPTEMBER
OCT.*
OCTOBER
NOV.*
NOVEMBER
DEC.*
DECEMBER
MON.*
MONDAY
TUE.*
TUESDAY
WED.*
WEDNESDAY
THU.*
THUR.*
THURSDAY
FRI.*
FRIDAY
SAT.*
SATURDAY
SUN.*
SUNDAY
19*
mid*

#TO
to *** and &RB_ to
to *** but &RB_ to
to

#TO_
_
to

#TOUCH
touch
touches
touched
touching

#V_
_
@V

#VBD
@VBD
@VBN

#VBD/VBG
@VBD
@VBN
@VBG

#WH
what
where
how
why
when
whether

#WHICH
which
that
who

#WHICH_
which
that
who
_

#WILL
will
would

APPENDIX 5. TITLES OF QUERY STATEMENTS

Appendix 5

Explanation of the symbols in the first column:

* indicates that a causal relation is central to the query

+ indicates that the query contains a causal relation but it is not central to the query

a indicates that the query was used in the model building step of the "ad hoc queries" experiment to determine the best set of weights to use for combining the scores from the different types of matching.

b indicates that the query was used in the model validation step of the "ad hoc queries" experiment to determine the retrieval effectiveness of the models developed in the model building step.

Within parenthesis, *total* indicates the total number of Wall Street Journal documents with relevance judgments from TREC-1 and TREC-2 for the query, and *rel* indicates the number of documents judged relevant.

*a	001 Antitrust Cases Pending (total:1142, rel:178)
	002 Acquisitions (total:1314, rel:244)
	003 Joint Ventures (total:1301, rel:366)
	004 Debt Rescheduling (total:998, rel:119)
+a	005 Dumping Charges (total:1044, rel:139)
	006 Third World Debt Relief (total:966, rel:204)
*b	007 U.S. Budget Deficit (total:1079, rel:160)
	008 Economic Projections (total:1378, rel:200)
	009 U.S. Congressional Positions on the SDI (total:304, rel:77)
*a	010 AIDS Treatments (total:691, rel:280)
	011 Space Program (total:688, rel:108)
*b	012 Water Pollution (total:1002, rel:131)
	013 Mitsubishi Heavy Industries Ltd. (total:936, rel:138)
+b	014 Drug Approval (total:981, rel:223)
*a	015 International Trade Liberalization Talks (total:333, rel:28)
	016 Marketing of Agrochemicals (total:1087, rel:132)
*b	017 Measures to Control Agrochemicals (total:853, rel:123)
	018 Japanese Stock Market Trends (total:976, rel:138)
*a	019 U.S. Protectionist Measures (total:354, rel:115)
	020 Patent Infringement Lawsuits (total:920, rel:336)
+a	021 Superconductors (total:713, rel:51)
*b	022 Counternarcotics (total:827, rel:171)
+b	023 Legal Repercussions of Agrochemical Use (total:921, rel:97)
+a	024 New Medical Technology (total:1078, rel:335))
*a	025 Aftermath of Chernobyl (total:607, rel:22)
*b	026 Tracking Influential Players in Multimedia (total:400, rel:73)
	027 Expert Systems and Neural Networks in Business or Manufacturing (total:155, rel:23)
	028 AT&T's Technical Efforts (total:546, rel:108)
	029 Foreign Installation of AT&T Communications Products (total:487, rel:92)
*a	030 OS/2 Problems (total:306, rel:57)
	031 Advantages of OS/2 (total:300, rel:24)

032 Who Outsources Computer Work to Whom (total:350, rel:19)

*b 033 Companies Capable of Producing Document Management (total:250, rel:19)

034 Entities Involved In Building ISDN Applications and (total:203, rel:13)

035 Alternatives to Postscript (total:280, rel:17)

036 How Rewritable Optical Disks Work (total:233, rel:3)

037 Identify SAA Components (total:275, rel:4)

*a 038 Impact of the "Religious Right" on U.S. Law (total:202, rel:57)

+b 039 Client-Server Plans and Expectations (total:243, rel:13)

*b 040 Analyses of Savings and Loan Failures (total:289, rel:150)

041 Computer or Communications Systems Upgrade (total:491, rel:20)

042 What is End User Computing and Who's Doing It (total:310, rel:108)

*a 043 U.S. Technology Policy (total:257, rel:60)

044 Staff Reductions at Computers and Communications (total:637, rel:127)

*b 045 What Makes CASE Succeed or Fail (total:298, rel:36)

046 Tracking Computer Virus Outbreaks (total:253, rel:11)

+a 047 Contracts for Computer Systems in Excess of $1 Million. (total:607, rel:178)

048 Purchasers of Modern Communications Equipment (total:462, rel:93)

049 Who's Working with Supercomputers (total:389, rel:66)

050 Potential Military Interest in Virtual Reality (total:319, rel:5)

*a 051 Airbus Subsidies (total:326, rel:58)

*b 052 South African Sanctions (total:230, rel:115)

053 Leveraged Buyouts (total:294, rel:140)

054 Satellite Launch Contracts (total:168, rel:38)

055 Insider Trading (total:481, rel:347)

*a 056 Prime (Lending) Rate Moves, Predictions (total:404, rel:397)

057 MCI (total:201, rel:71)

+b 058 Rail Strikes (total:191, rel:25)

* 059 Weather Related Fatalities (total:21, rel:0)

*a 060 Merit-Pay vs. Seniority (total:166, rel:23)

061 Israeli Role in Iran-Contra Affair (total:264, rel:106)

+a 062 Military Coups D'etat (total:112, rel:28)

+b 063 Machine Translation (total:6, rel:1)

*b 064 Hostage-Taking (total:150, rel:29)

065 Information Retrieval Systems (total:14, rel:3)

*a 066 Natural Language Processing (total:7, rel:1)

*b 067 Politically Motivated Civil Disturbances (total:87, rel:36)

*a 068 Health Hazards from Fine-Diameter Fibers (total:145, rel:26)

*b 069 Attempts to Revive the SALT II Treaty (total:429, rel:28)

+a 070 Surrogate Motherhood (total:317, rel:11)

+b 071 Border Incursions (total:88, rel:29)

+a 072 Demographic Shifts in the U.S. (total:292, rel:25)

+b 073 Demographic Shifts across National Boundaries (total:318, rel:38)

074 Conflicting Policy (total:350, rel:32)

*a 075 Automation (total:191, rel:24)

076 U.S. Constitution - Original Intent (total:146, rel:62)
+a 077 Poaching (total:209, rel:5)
+b 078 Greenpeace (total:91, rel:1)
079 FRG Political Party Positions (total:293, rel:44)
080 1988 Presidential Candidates Platforms (total:499, rel:185)
*b 081 Financial Crunch for Televangelists in the Wake of the PTL (total:451, rel:9)
+a 082 Genetic Engineering (total:133, rel:106)
*a 083 Measures to Protect the Atmosphere (total:62, rel:51)
084 Alternative/renewable Energy Plant & Equipment Installation (total:93, rel:51)
*b 085 Official Corruption (total:212, rel:130)
086 Bank Failures (total:230, rel:89)
087 Criminal Actions Against Officers of Failed Financial Institutions (total:446, rel:30)
088 Crude Oil Price Trends (total:257, rel:61)
089 "Downstream" Investments by OPEC Member States (total:315, rel:39)
+b 090 Data on Proven Reserves of Oil & Natural Gas (total:227, rel:84)
091 U.S. Army Acquisition of Advanced Weapons Systems (total:507, rel:15)
092 International Military Equipment Sales (total:486, rel:15)
093 What Backing Does the National Rifle Association (total:320, rel:9)
094 Computer-aided Crime (total:314, rel:18)
095 Computer-aided Crime Detection (total:116, rel:18)
096 Computer-Aided Medical Diagnosis (total:67, rel:10)
097 Fiber Optics Applications (total:22, rel:7)
*a 098 Fiber Optics Equipment Manufacturers (total:29, rel:18)
099 Iran-Contra Affair (total:304, rel:90)
100 Controlling the Transfer of High Technology (total:178, rel:40)
+a 101 Design of the "Star Wars" Anti-missile Defense System (total:337, rel:7)
102 Laser Research Applicable to the U.S.'s Strategic Defense (total:410, rel:7)
*b 103 Welfare Reform (total:372, rel:25)
*a 104 Catastrophic Health Insurance (total:394, rel:25)
*b 105 "Black Monday" (total:683, rel:26)
*a 106 U.S. Control of Insider Trading (total:757, rel:146)
107 Japanese Regulation of Insider Trading (total:900, rel:65)
*b 108 Japanese Protectionist Measures (total:818, rel:189)
109 Find Innovative Companies (total:413, rel:219)
*a 110 Black Resistance Against the South African Government (total:295, rel:150)
111 Nuclear Proliferation (total:362, rel:124)
112 Funding Biotechnology (total:912, rel:334)
113 New Space Satellite Applications (total:434, rel:72)
114 Non-commercial Satellite Launches (total:375, rel:20)
*b 115 Impact of the 1986 Immigration Law (total:209, rel:85)
+b 116 Generic Drug Substitutions (total:458, rel:28)
117 Capacity of the U.S. Cellular Telephone Network (total:469,

rel:106)
118 International Terrorists (total:341, rel:87)
119 Actions Against International Terrorists (total:400, rel:84)
*a 120 Economic Impact of International Terrorism (total:572, rel:48)
*b 121 Death from Cancer (total:466, rel:2)
+a 122 RDT&E of New Cancer Fighting Drugs (total:589, rel:86)
*a 123 Research into & Control of Carcinogens (total:380, rel:106)
*b 124 Alternatives to Traditional Cancer Therapies (total:516, rel:77)
*a 125 Anti-smoking Actions by Government (total:397, rel:76)
+b 126 Medical Ethics and Modern Technology (total:427, rel:57)
127 U.S.-U.S.S.R. Arms Control Agreements (total:374, rel:68)
128 Privatization of State Assets (total:870, rel:296)
129 Soviet Spying on the U.S. (total:428, rel:49)
*b 130 Jewish Emigration and U.S.-USSR Relations (total:317, rel:64)
+a 131 McDonnell Douglas Contracts for Military Aircraft (total:659, rel:22)
132 "Stealth" Aircraft (total:361, rel:64)
+b 133 Hubble Space Telescope (total:258, rel:29)
134 The Human Genome Project (total:243, rel:23)
*a 135 Possible Contributions of Gene Mapping to Medicine (total:268, rel:156)
136 Diversification by Pacific Telesis (total:466, rel:121)
+a 137 Expansion in the U.S. Theme Park Industry (total:513, rel:111)
138 Iranian Support for Lebanese Hostage-takers (total:264, rel:20)
*b 139 Iran's Islamic Revolution - Domestic and Foreign Social (total:332, rel:10)
*a 140 Political Impact of Islamic Fundamentalism (total:448, rel:6)
*b 141 Japan's Handling of its Trade Surplus with the U.S. (total:779, rel:16)
*a 142 Impact of Government Regulated Grain Farming on International (total:627, rel:338)
*b 143 Why Protect U.S. Farmers? (total:502, rel:135)
144 Management Problems at the United Nations (total:476, rel:7)
*a 145 Influence of the "Pro-Israel Lobby" (total:527, rel:64)
*b 146 Negotiating an End to the Nicaraguan Civil War (total:311, rel:68)
147 Productivity Trends in the U.S. Economy (total:774, rel:209)
148 Conflict in the Horn of Africa (total:227, rel:77)
+b 149 Industrial Espionage (total:442, rel:98)
150 U.S. Political Campaign Financing (total:491, rel:254)

APPENDIX 6. RESULTS OF PRELIMINARY RETRIEVAL EXPERIMENTS

Appendix 6

A. Retrieval results for *keyword matching* using various weighting schemes
B. Retrieval results for *Roget code matching* using various weighting schemes
C. Retrieval results for *Roget code matching* with each Roget code weighted inversely by the number of codes assigned to a word. (Explained in Chapter 5 Section 5.3.2.4)

The following retrieval measures are used in the tables:

- The precision for 11 recall levels
- The 11-point recall-precision average
- The 3-point recall-precision average
- The normalized recall
- The normalized precision

Results are averaged over 39 query statements.

A. Retrieval Results for *Keyword Matching* using Various Weighting Schemes

Words are stemmed to an entry in Longman Dictionary of Contemporary English.

```
    Weighting scheme
        1      2      3      4      5      6      7      8      9      10     11     12
  13     14     15     16     17
        Doc wt:

        bin    tf ---------------->|tf-cos ------------>|tfidf
-------------------------->|tfidf-cos ----------------------->|
        Query wt:
        bin    bin    ntf    tf     bin    ntf    tf     bin    ntf    tf     ntfidf tfidf
 bin    ntf    tf     ntfidf tfidf

Precision at 11 recall levels
0%
|0.7469|0.6403|0.6572|0.6793|0.8760|0.8674|0.8131|0.6957|0.7368|0.7408|0.7406|0.7668|0.851
5|0.8630|0.8188|0.8431|0.8445|
10%
|0.5794|0.4926|0.5018|0.5271|0.7422|0.7618|0.6987|0.5917|0.6193|0.6232|0.6352|0.6657|0.737
6|0.7457|0.7014|0.7116|0.7156|
20%
|0.5068|0.4206|0.4258|0.4566|0.6473|0.6553|0.5930|0.5043|0.5176|0.5370|0.5556|0.5662|0.620
4|0.6392|0.5804|0.6472|0.6077|
30%
|0.4564|0.3933|0.3965|0.4286|0.5862|0.5953|0.5499|0.4639|0.4735|0.5038|0.5206|0.5249|0.571
9|0.5800|0.5538|0.5881|0.5626|
40%
|0.4315|0.3674|0.3738|0.4061|0.5261|0.5320|0.5163|0.4412|0.4513|0.4666|0.4769|0.4850|0.529
6|0.5526|0.5283|0.5547|0.5376|
50%
|0.4048|0.3563|0.3622|0.3905|0.4807|0.4898|0.4855|0.4212|0.4269|0.4390|0.4585|0.4671|0.489
7|0.5033|0.4952|0.5197|0.4994|
60%
|0.3841|0.3441|0.3454|0.3689|0.4425|0.4535|0.4468|0.3938|0.4022|0.4131|0.4274|0.4391|0.451
```

```
9|0.4519|0.4477|0.4736|0.4682|
70%
|0.3587|0.3333|0.3347|0.3523|0.4163|0.4214|0.4110|0.3726|0.3759|0.3850|0.4019|0.4096|0.417
2|0.4199|0.4142|0.4313|0.4267|
80%
|0.3414|0.3247|0.3260|0.3413|0.3884|0.3927|0.3818|0.3598|0.3620|0.3680|0.3796|0.3858|0.392
2|0.3947|0.3835|0.4035|0.3997|
90%
|0.3184|0.3091|0.3116|0.3280|0.3523|0.3523|0.3436|0.3389|0.3417|0.3499|0.3560|0.3658|0.357
1|0.3599|0.3539|0.3704|0.3639|
100%
|0.2694|0.2649|0.2663|0.2825|0.2888|0.2902|0.2780|0.2870|0.2889|0.2927|0.2955|0.3011|0.293
9|0.2969|0.2865|0.3043|0.2949|

Average precision over the 11 recall levels
mean
|0.4362|0.3861|0.3910|0.4146|0.5224|0.5283|0.5016|0.4427|0.4542|0.4654|0.4771|0.4888|0.519
3|0.5279|0.5058|0.5316|0.5201|
std
dev|0.2414|0.2443|0.2451|0.2639|0.2411|0.2419|0.2301|0.2615|0.2576|0.2540|0.2550|0.2504|0.
2421|0.2387|0.2285|0.2412|0.2232|

Normalized recall
mean
|0.6969|0.6410|0.6474|0.6672|0.7532|0.7583|0.7524|0.6945|0.7013|0.7148|0.7313|0.7415|0.758
7|0.7639|0.7610|0.7790|0.7752|
std
dev|0.1171|0.1270|0.1252|0.1286|0.1100|0.1041|0.1021|0.1374|0.1349|0.1333|0.1323|0.1318|0.
1129|0.1092|0.1057|0.1098|0.1040|

Normalized precision
mean
|0.5652|0.4888|0.4951|0.5205|0.6497|0.6607|0.6452|0.5605|0.5713|0.5884|0.6053|0.6188|0.651
7|0.6603|0.6482|0.6651|0.6583|
std
dev|0.1849|0.1836|0.1849|0.1793|0.1635|0.1595|0.1427|0.1870|0.1895|0.1885|0.1881|0.1877|0.
```

```
1782|0.1741|0.1536|0.1646|0.1569|
```

B. Retrieval Results for *Roget Code Matching* using Various Weighting Schemes

Words in text are replaced with Roget category codes.

```
    Weighting scheme
        1      2      3      4      5      6      7      8      9      10     11     12
  13     14     15     16     17
        Doc wt:
        bin    tf ---------------->|tf-cos ------------>|tfidf
--------------------------->|tfidf-cos ----------------------->|
        Query wt:
        bin    bin    ntf    tf     bin    ntf    tf     bin    ntf    tf     ntfidf tfidf
 bin    ntf    tf     ntfidf tfidf

Precision at 11 recall levels
0%
|0.6024|0.4076|0.4120|0.4397|0.7113|0.7403|0.8172|0.4969|0.5039|0.5246|0.6209|0.6809|0.689
6|0.7049|0.7294|0.7447|0.7255|
10%
|0.4091|0.3216|0.3258|0.3365|0.5762|0.6115|0.6325|0.3789|0.3849|0.4075|0.4982|0.5398|0.539
1|0.5599|0.5779|0.6161|0.6046|
20%
|0.3449|0.3083|0.3119|0.3205|0.4935|0.5140|0.5263|0.3392|0.3420|0.3593|0.4467|0.4499|0.480
0|0.4847|0.4737|0.5475|0.5326|
30%
|0.3293|0.2997|0.3016|0.3106|0.4537|0.4650|0.4857|0.3241|0.3277|0.3449|0.4109|0.4237|0.431
0|0.4382|0.4367|0.5013|0.4983|
40%
|0.3161|0.2976|0.2995|0.3086|0.4091|0.4219|0.4353|0.3178|0.3200|0.3318|0.3828|0.4026|0.378
4|0.3911|0.4072|0.4752|0.4718|
50%
|0.3060|0.2959|0.2975|0.3051|0.3886|0.4016|0.4115|0.3120|0.3141|0.3225|0.3694|0.3857|0.357
1|0.3657|0.3853|0.4327|0.4452|
60%
|0.2984|0.2943|0.2957|0.3013|0.3637|0.3749|0.3848|0.3081|0.3088|0.3168|0.3423|0.3592|0.342
2|0.3445|0.3534|0.3919|0.4009|
70%
|0.2936|0.2909|0.2921|0.2964|0.3411|0.3500|0.3581|0.3033|0.3048|0.3098|0.3304|0.3468|0.328
8|0.3316|0.3344|0.3669|0.3738|
80%
|0.2886|0.2878|0.2884|0.2925|0.3268|0.3327|0.3389|0.2978|0.2984|0.3039|0.3207|0.3399|0.311
3|0.3152|0.3208|0.3457|0.3561|
90%
|0.2827|0.2834|0.2836|0.2859|0.2979|0.3038|0.3137|0.2905|0.2915|0.2969|0.3070|0.3254|0.299
5|0.3012|0.3034|0.3243|0.3288|
100%
|0.2506|0.2526|0.2530|0.2546|0.2544|0.2561|0.2573|0.2572|0.2576|0.2599|0.2628|0.2793|0.259
1|0.2600|0.2595|0.2699|0.2737|

Average precision over the 11 recall levels
mean
|0.3383|0.3036|0.3056|0.3138|0.4197|0.4338|0.4510|0.3296|0.3322|0.3434|0.3902|0.4121|0.401
5|0.4088|0.4165|0.4560|0.4556|
std
dev|0.2377|0.2438|0.2440|0.2477|0.2445|0.2440|0.2377|0.2460|0.2461|0.2450|0.2472|0.2540|0.
2436|0.2436|0.2437|0.2464|0.2460|

Normalized recall
mean
|0.5664|0.5207|0.5248|0.5417|0.6637|0.6777|0.6881|0.5657|0.5701|0.5868|0.6438|0.6565|0.633
6|0.6423|0.6548|0.7103|0.7166|
std
dev|0.1191|0.1175|0.1173|0.1182|0.1062|0.1060|0.1160|0.1225|0.1226|0.1236|0.1296|0.1374|0.
1123|0.1121|0.1183|0.1292|0.1287|

Normalized precision
mean
|0.4272|0.3568|0.3582|0.3728|0.5362|0.5536|0.5678|0.4022|0.4064|0.4240|0.4903|0.5109|0.503
1|0.5105|0.5213|0.5752|0.5790|
std
dev|0.1650|0.1576|0.1575|0.1591|0.1802|0.1804|0.1845|0.1695|0.1676|0.1662|0.1774|0.1659|0.
1787|0.1801|0.1793|0.1954|0.1823|
```

C. Retrieval Results for *Roget Code Matching* with each Roget Code Weighted Inversely by the Number of Codes Assigned to a Word.

Roget codes are weighted inversely by the number of Roget codes assigned to a word. This is explained in Chapter 5 (Section 5.3.2.4).

```
    Weighting scheme
        1      2      3      4      5      6      7      8      9      10     11     12
  13     14     15     16     17
        Doc wt:
        bin    tf ---------------->|tf-cos ------------>|tfidf
--------------------------->|tfidf-cos ----------------------->|
        Query wt:
        bin    bin    ntf    tf     bin    ntf    tf     bin    ntf    tf     ntfidf tfidf
 bin    ntf    tf     ntfidf tfidf
```

<u>Precision at 11 recall levels</u>

```
0%
|0.6024|0.4511|0.4647|0.5396|0.7332|0.7503|0.8347|0.5932|0.6186|0.6369|0.6541|0.6326|0.673
7|0.7107|0.7010|0.7048|0.7083|
10%
|0.4091|0.3455|0.3622|0.4341|0.6017|0.6452|0.6339|0.4488|0.4558|0.5230|0.5493|0.5538|0.528
3|0.5730|0.5644|0.5929|0.5759|
20%
|0.3449|0.3188|0.3248|0.3898|0.5295|0.5731|0.5669|0.4062|0.4248|0.4569|0.4856|0.4890|0.474
7|0.4921|0.5002|0.5271|0.5137|
30%
|0.3293|0.3144|0.3201|0.3754|0.4887|0.5270|0.5133|0.3826|0.4016|0.4304|0.4589|0.4665|0.452
1|0.4623|0.4665|0.4899|0.4836|
40%
|0.3161|0.3108|0.3144|0.3576|0.4480|0.4756|0.4691|0.3610|0.3743|0.4077|0.4352|0.4428|0.408
0|0.4253|0.4486|0.4694|0.4681|
50%
|0.3060|0.3051|0.3091|0.3492|0.4210|0.4410|0.4500|0.3425|0.3539|0.3908|0.4167|0.4288|0.390
8|0.4118|0.4365|0.4483|0.4506|
60%
```

```
|0.2984|0.3024|0.3059|0.3397|0.3712|0.3889|0.4096|0.3304|0.3346|0.3621|0.3843|0.4065|0.359
2|0.3677|0.4058|0.4141|0.4182|
70%
|0.2936|0.2986|0.3019|0.3308|0.3535|0.3671|0.3799|0.3231|0.3264|0.3492|0.3710|0.3884|0.346
5|0.3522|0.3766|0.3927|0.3992|
80%
|0.2886|0.2941|0.2962|0.3225|0.3392|0.3497|0.3574|0.3154|0.3192|0.3345|0.3462|0.3664|0.328
1|0.3347|0.3547|0.3626|0.3726|
90%
|0.2827|0.2877|0.2898|0.3121|0.3188|0.3251|0.3313|0.3006|0.3045|0.3181|0.3206|0.3360|0.311
4|0.3150|0.3256|0.3313|0.3422|
100%
|0.2506|0.2553|0.2560|0.2742|0.2709|0.2745|0.2739|0.2598|0.2616|0.2698|0.2678|0.2725|0.263
9|0.2656|0.2715|0.2706|0.2752|

Average precision over the 11 recall levels
mean
|0.3383|0.3167|0.3223|0.3659|0.4432|0.4652|0.4746|0.3694|0.3796|0.4072|0.4263|0.4348|0.412
4|0.4282|0.4410|0.4549|0.4552|
std
dev|0.2377|0.2463|0.2472|0.2675|0.2616|0.2603|0.2358|0.2479|0.2462|0.2409|0.2527|0.2573|0.
2478|0.2425|0.2463|0.2514|0.2515|

Normalized recall
mean
|0.5664|0.5482|0.5586|0.6107|0.6686|0.6905|0.7171|0.6189|0.6309|0.6661|0.6866|0.7024|0.667
2|0.6792|0.7038|0.7167|0.7233|
std
dev|0.1191|0.1226|0.1228|0.1377|0.1306|0.1310|0.1350|0.1310|0.1325|0.1421|0.1509|0.1523|0.
1251|0.1280|0.1366|0.1425|0.1388|

Normalized precision
mean
|0.4272|0.3832|0.3933|0.4544|0.5482|0.5800|0.6048|0.4663|0.4757|0.5166|0.5414|0.5545|0.524
3|0.5381|0.5583|0.5760|0.5766|
std
```

```
dev|0.1650|0.1619|0.1628|0.1751|0.1909|0.1914|0.1613|0.1756|0.1765|0.1864|0.2058|0.2129|0.
1834|0.1842|0.1855|0.2045|0.1988|
```

BIBLIOGRAPHY

Bibliography

Ajzen, Icek. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology*, **35**(5), 303-314.

Ajzen, Icek, & Fishbein, Martin. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin*, **82**(2), 261-277.

Alloy, Lauren B., & Tabachnik, Naomi. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, **91**(1), 112-149.

Altenberg, Bengt. (1984). Causal linking in spoken and written English. *Studia Linguistica*, **38**(1), 20-69.

Asher, R.E. (Ed.). (1994). *The encyclopedia of language and linguistics*. Oxford: Pergamon Press.

Au, Terry Kit-Fong. (1986). A verb is worth a thousand words: The cause and consequences of interpersonal events implicit in language. *Journal of Memory and Language*, **25**(1), 104-122.

Austin, Derek. (1984). *PRECIS: A manual of concept analysis and subject indexing.* (2nd ed.). London: British Library, Bibliographic Services Division.

Belkin, N.J., Oddy, R.N., & Brooks, H.M. (1982a). ASK for information retrieval: Part I, Background and theory. *Journal of Documentation*, **38**(2), 61-71.

Belkin, N.J., Oddy, R.N., & Brooks, H.M. (1982b). ASK for information retrieval: Part II, Results of a design study. *Journal of Documentation*, **38**(3), 145-164.

Berrut, Catherine. (1990). Indexing medical reports: the RIME approach. *Information Processing & Management*, **26**(1), 93-109.

Bozsahin, H. Cem, & Findler, Nicholas V. (1992). Memory-based hypothesis formation: Heuristic learning of commonsense causal relations from text. *Cognitive Science*, **16**(4), 431-454

Brown, Roger, & Fish, Deborah. (1983). The psychological causality implicit in language. *Cognition*, **14**(3), 237-273.

Caramazza, Alfonso, Grober, Ellen, Garvey, Catherine, & Yates, Jack. (1977). Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior*, **16**(5), 601-609.

Carrier, Jill, & Randall, Janet H. (1992). The argument structure and syntactic structure of resultatives. *Linguistic Inquiry*, **23**(2), 173-234.

Chan, C.W. (1995). The development and application of a knowledge modelling technique. *Journal of Experimental and Theoretical Artificial Intelligence*, **7**(2), 217-236.

Cheng, Patricia W., & Holyoak, Keith J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, **17**(4), 391-416.

Cheng, Patricia W., & Novick, Laura R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, **58**(4), 545-567.

Cheng, Patricia W., & Novick, Laura R. (1992). Covariation in natural causal induction. *Psychological Review*, **99**(2), 365-382.

Chorafas, Dimitris N. (1990). *Knowledge engineering: Knowledge acquisition, knowledge representation, the role of the knowledge engineer, and domains fertile* *for AI implementation*. New York: Van Nostrand Reinhold.

Corrigan, Roberta. (1988). Who dun it? The influence of actor-patient animacy and type of verb in the making of causal attributions. *Journal of Memory and Language*, **27**(4), 447-465.

Cresswell, M.J. (1981). Adverbs of causation. In H.-J. Eikmeyer & H. Rieser (Eds.), *Words, worlds, and contexts: New approaches in word semantics* (pp. 21-37). Berlin: Walter de Gruyter.

Croft, W. Bruce. (1986). Boolean queries and term dependencies in probabilistic retrieval models. *Journal of the American Society for Information Science*, **37**(2), 71-77.

Croft, W. Bruce, Turtle, Howard R., & Lewis, David D. (1991). The Use of Phrases and Structured Queries in Information Retrieval. In A. Bookstein, Y. Chiaramella, G. Salton, & V.V. Raghavan (Eds.), *SIGIR '91: Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (pp. 32-45). New York: ACM Press.

Cullingford, R.E. (1978). *Script applications: Computer understanding of newspaper stories* (Tech. Rep. No. 116.). New Haven, CT: Yale University, Department of Computer Science.

Cummins, Denise D., Lubart, Todd, Alksnis, Olaf, & Rist, Robert. (1991). Conditional reasoning and causation. *Memory & Cognition*, **19**(3), 274-282.

Cunningham, John D., & Kelley, Harold H. (1975). Causal attributions for interpersonal events of varying magnitude. *Journal of Personality*, **43**(1), 74-93.

Dakin, Julian. (1970). Explanations. *Journal of Linguistics*, **6**(2), 199-214.

Davidson, Donald. (1980). Causal relations. In D. Davidson, *Essays on actions and events* (pp. 149-162). Oxford: Oxford University Press.

Dillon, Martin, & Gray, Ann S. (1983). FASIT: A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, **34**(2), 99-108.

Downing, Cathryn J., Sternberg, Robert J., & Ross, Brian H. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. *Journal of Experimental Psychology: General*, **114**(2), 239-263.

Einhorn, Hillel J., & Hogarth, Robin M. (1986). Judging probable cause. *Psychological Bulletin*, **99**(1), 3-19.

Emmet, Dorothy. (1985). *The Effectiveness of Causes*. Albany, NY: State University of New York Press.

Fagan, Joel L. (1989). The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, **40**(2), 115-132

Fales, Evan. (1990). *Causation and universals*. London: Routledge.

Farradane, J.E.L. (1950). A scientific theory of classification and indexing and its practical applications. *Journal of Documentation*, **6**(2), 83-99.

Farradane, J.E.L. (1952). A scientific theory of classification and indexing: further considerations. *Journal of Documentation*, **8**(2), 73-92.

Farradane, J.E.L. (1967). Concept organization for information retrieval. *Information Storage and Retrieval*, **3**(4), 297-314.

Feigenbaum, E.A. (1984). Knowledge engineering: The applied side of artificial intelligence. *Annals of the New York Acedemy of Sciences*, **246**, 91-107.

Fillmore, Charles J. (1968). The case for case. In E. Bach & R. T. Harms (Eds.), *Universals in Linguistic Theory* (pp.1-88). New York : Holt, Rinehart and Winston.

Fletcher, Charles R., & Bloom, Charles P. (1988). Causal reasoning in the comprehension of simple narrative texts. *Journal of Memory and Language*, **27**(3), 235-244.

Fox, Edward A. (1983). *Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts* (Report No. TR83-561). Ithaca, NY: Department of Computer Science, Cornell University.

Gaines, B.R., & Shaw, M.L.G. (1991). Foundations of knowledge acquisition. In H. Motoda, et al. (Eds.), *Knowledge acquisition for knowledge-based systems* (pp.3-24). Amsterdam: IOS.

Gardin, Jean-Claude. (1965). *SYNTOL.* New Brunswick, NJ: Graduate School of Library Service, Rutgers, The State University.

Garvey, Catherine, & Caramazza, Alfonso. (1974). Implicit causality in verbs. *Linguistic Inquiry*, **5**(3), 459-464.

Garvey, Catherine, Caramazza, Alfonso, & Yates, Jack. (1974/1975). Factors influencing assignment of pronoun antecedents. *Cognition*, **3**(3), 227-243.

Gay, L.S., & Croft, W.B. (1990). Interpreting nominal compounds for information retrieval. *Information Processing & Management*, **26**(1), 21-38.

Goldberg, Adele E. (1991). A semantic account of resultatives. *Linguistic Analysis*, **21**(1-2), 66-96.

Gonsalves, Renison J. (1986). A decompositional analysis of causative verbs. *CUNYforum*, **12**, 31-65.

Greenbaum, S. (1969). *Studies in english adverbial usage.* London: Longman.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hansen, Ronald D., & Donoghue, James M. (1977). The power of consensus: Information derived from one's own and others' behavior. *Journal of Personality and Social Psychology*, **35**(5), 294-302.

Harman, Donna. (1993a). Overview of the First Text REtrieval Conference (TREC-1). In D.K. Harman (Ed.), *The First Text REtrieval Conference (TREC-1)* (NIST Special Publication 500-207, pp. 1-20). Gaithersburg, MD: National Institute of Standards and Technology.

Harman, Donna. (1993b). The DARPA TIPSTER project. *SIGIR Forum*, **26**(2), 26-28.

Harman, Donna (Ed.). (1993c). *The First Text REtrieval Conference (TREC-1)* (NIST Special Publication 500-207). Gaithersburg, MD: National Institute of Standards and Technology.

Harman, Donna. (1994a) Overview of the Second Text REtrieval Conference (TREC-2). In D.K. Harman (Ed.), *The Second Text REtrieval Conference (TREC-2)* (NIST Special Publication 500-215, pp. 1-20). Gaithersburg, MD: National Institute of Standards and Technology.

Harman, Donna (Ed.). (1994b). *The Second Text REtrieval Conference (TREC-2)*

(NIST Special Publication 500-215). Gaithersburg, MD: National Institute of Standards and Technology.

Harman, Donna (Ed.). (1995). *Overview of the Third Text REtrieval Conference (TREC-3)* (NIST Special Publication 500-225). Gaithersburg, MD: National Institute of Standards and Technology.

Harris, Roy. (1987). *Reading Saussure: A critical commentary on the Cours de linguistique generale*. La Salle, Ill.: Open Court.

Hesslow, Germund. (1988). The problem of causal selection. In D.J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11-32). Washington Square, NY: New York University Press.

Hewstone, Miles, & Jaspars, Jos. (1983). A re-examination of the roles of consensus, consistency and distinctiveness: Kelley's cube revisited. *British Journal of Social Psychology*, **22**(1), 41-50.

Hilton, Denis J. (1988). Logic and causal attribution. In D.J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 33-65). Washington Square, NY: New York University Press.

Hilton, Denis J., & Jaspars, Joseph M.F. (1987). The explanation of occurrences and non-occurrences: A test of the inductive logic model of causal attribution. *British Journal of Social Psychology*, **26**, 189-201.

Hilton, Denis J., Jaspars, Joseph M.F., & Clarke, David D. (1990). Pragmatic conditional reasoning: Context and content effects on the interpretation of causal assertions. *Journal of pragmatics*, **14**(5), 791-812.

Hilton, Denis J., & Slugoski, B.R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, **93**, 75-88.

Hoekstra, Teun. (1988). Small clause results. *Lingua*, **74**(2/3), 101-139.

Hosmer, David W., Jr., & Lemeshow, Stanley. (1989). *Applied logistic regression*. New York: Wiley.

Hume, David. (1911). *A treatise of human nature*. London: J.M. Dent & Sons. (Original work published 1738).

Hume, David. (1965). *An abstract of a treatise of human nature*. Hamden, Connecticut: Archon Books. (Original work published 1740).

Jackson, Peter. (1989). Applications of nonmonotonic logic to diagnosis. *Knowledge Engineering Review*, **4**(2), 97-117.

Jackson, Peter. (1990). *Introduction to expert systems*. (2nd ed.). Wokingham, England: Addison-Wesley.

Jaspars, Jos. (1983). The process of attribution in common-sense. In M. Hewstone (Ed.), *Attribution theory: social and functional extensions* (pp. 28-44). Oxford: Basil Blackwell.

Jaspars, Jos, Hewstone, Miles, & Fincham, Frank D. (1983). Attribution theory and research: The state of the art. In J. Jaspars, F.D. Fincham, & M. Hewstone (Eds.), *Attribution theory and research: Conceptual, developmental and social dimensions* (pp. 3-36). London: Academic Press.

Jones, Leslie P., deBessonet, Cary, & Kundu, Sukhamay. (1988). ALLOY: An amalgamation of expert, linguistic and statistical indexing methods. In Y. Chiaramella (Ed.), *11th International Conference on Research & Development* *in Information Retrieval* (pp. 191-199). New York: ACM.

Joskowsicz, Leo, Ksiezyk, Tomasz, & Grishman, Ralph. (1989). Deep domain models for discourse analysis. In H.J. Antonisse, J.W. Benolt, & B.G. Silverman (Eds.), *The Annual AI Systems in Government Conference* (pp. 195-200). Silver Spring, MD: IEEE Computer Society.

Kaplan, Randy M., & Berry-Rogghe, Genevieve. (1991). Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition*, **3**(3), 317-337.

Kassin, Saul M. (1979). Consensus information, prediction, and causal attribution: A review of the literature and issues. *Journal of Personality and Social Psychology*, **37**(11), 1966-1981.

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

Keenan, Janice M., Baillet, Susan D., & Brown, Polly. (1984). The effects of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior*, **23**(2), 115-126.

Kelley, Harold H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, 192-238. (Current Theory and Research in Motivation, v. 15)

Kelley, Harold H. (1973). The process of causal attribution. *American Psychologist*, **28**(2), 107-128.

Kishore, Jugal. (1986). *Colon Classification: Enumerated & expanded schedules along with theoretical formulations*. New Delhi: Ess Ess Publications.

Kontos, John, & Sidiropoulou, Maria. (1991). On the acquisition of causal knowledge from scientific texts with attribute grammars. *Expert Systems for Information Management*, **4**(1), 31-48.

Kovalyova, L.M. (1979). On predicates *kill* and *cause to die*. *Zeitschrift fur Anglistik und Amerikanistik*, **27**(2), 146-153.

Kovalyova, L.M. (1988). Presuppositions on concrete nouns in causation constructions. *Zeitschrift fur Anglistik und Amerikanistik*, **36**(3), 235-240.

Kulik, James A., & Taylor, Shelley E. (1980). Premature consensus on consensus? Effects of sample-based versus self-based consensus information. *Journal of Personality and Social Psychology*, **38**(6), 871-878.

Kun, Anna, & Weiner, Bernard. (1973). Necessary and sufficient causal schemata for success and failure. *Journal of Research in Personality*, **7**(3), 197-207.

Lancaster, F.W. (1986). *Vocabulary control for information retrieval.* (2nd ed.). Arlington, VA: Information Resources Press.

Lebowitz, Michael. (1980). *Generalization and memory in an integrated understanding system* (Tech. Rep. No. 186.). New Haven, CT: Yale University, Department of Computer Science.

Levy, Francis. (1967). On the relative nature of relational factors in classifications. *Information Storage & Retrieval*, **3**(4), 315-329.

Liddy, Elizabeth D., & Myaeng, Sung H. (1993). DR-LINK's linguistic-conceptual approach to document detection. In D.K. Harman (Ed.), *The First Text REtrieval Conference (TREC-1)* (NIST Special Publication 500-207, pp. 1-20). Gaithersburg, MD: National Institute of Standards and Technology.

Liddy, Elizabeth D., & Myaeng, Sung H. (1994). DR-LINK: A system update for TREC-2. In D.K. Harman (Ed.), *The Second Text REtrieval Conference (TREC-2)* (NIST Special Publication 500-215, pp. 85-100). Gaithersburg, MD: National Institute of Standards and Technology.

Longman dictionary of contemporary English. (1987). 2nd ed. Harlow, Essex: Longman.

Lopez-Lopez, Aurelio. (1995). Beyond topicality: Exploiting the metadiscourse of abstracts to retrieve documents with analogous features. Unpublished doctoral dissertation, Syracuse University, New York.

Lu, Xin. (1990). An application of case relations to document retrieval (Doctoral dissertation, University of Western Ontario, 1990). *Dissertation Abstracts International*, **52-10**, 3464A.

Lyons, John. (1977). *Semantics. Vol. 2.* Cambridge: Cambridge University Press.

Mackie, J.L. (1980). *The cement of the universe: A study of causation*. Oxford: Oxford University Press. (Same as the 1974 edition with corrections and an additional preface.)

Marik, V., & Vlcek, T. (1992). Some aspects of knowledge engineering. In V.Marik, O. Stepankova & R. Trappl (Eds.), *Advanced Topics in Artificial Intelligence: International Summer School Proceedings* (pp. 316-337). Berlin: Springer-Verlag.

Mauldin, Michael L. (1991). Retrieval performance in FERRET: A conceptual information retrieval system. In A. Bookstein, Y. Chiaramella, G. Salton, & V.V. Raghavan (Eds.), *SIGIR '91: Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (pp. 347-355). New York: ACM Press.

McArthur, L.A. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology*, **22**, 171-193.

Meteer, M., Schwartz, R., & Weischedel, R. (1991). POST: Using probabilities in language processing. In *IJCAI-91: Proceedings of the Twelfth International Conference on Artificial Intelligence* (pp. 960-965). San Mateo, CA: Distributed by Morgan Kaufman Publishers.

Metzler, Douglas P., & Haas, Stephanie W. (1989). The Constituent Object Parser: Syntactic structure matching for information retrieval. *ACM Transactions on Information Systems*, **7**(3), 292-316.

Metzler, Douglas P., Haas, Stephanie W., Cosic, Cynthia L., & Weise, Charlotte A. (1990). Conjunction, ellipsis, and other discontinuous constituents in the constituent object parser. *Information Processing & Management*, **26**(1), 53-71.

Metzler, Douglas P., Haas, Stephanie W., Cosic, Cynthia L., & Wheeler, Leslie H. (1989). Constituent object parsing for information retrieval and similar text processing problems. *Journal of the American Society for Information Science*, **40**(6), 398-423.

Mill, John Stuart. (1973). A system of logic: Ratiocinative and inductive. Book

III: Of induction. In J.M. Robson (Ed.), *Collected works of John Stuart Mill* (vol. VII). Toronto: University of Toronto Press. (Original work published in 1872).

Mooney, Raymond J. (1990). Learning plan schemata from observation: Explanation-based learning for plan recognition. *Cognitive Science*, **14**(4), 483-509.

Myaeng, Sung H., & Liddy, Elizabeth D. (1993). Information retrieval with semantic representation of texts. In *Proceedings of the 2nd Annual Symposium on Document Analysis and Information Retrieval* (pp. 201-215).

Myaeng, Sung H., Khoo, Christopher, & Li, Ming. (1994). Linguistic processing of text for a large-scale conceptual information retrieval system. In Tepfenhart, W.M., Dick, J.P., & Sowa, J.F. (Eds.), *Conceptual Structures: Current Practices: Second International Conference on Conceptual Structures, ICCS '94* (pp. 69-83). Berlin: Springer-Verlag.

Newman, Andrew. (1988). The causal relation and its terms. *Mind*, **xcvii**(388), 529-550.

Nisbett, Richard E., & Borgida, Eugene. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, **32**(5), 932-943.

Nishida, Fujio, & Takamatsu, Shinobu. (1982). Structured-information extraction from patent-claim sentences. *Information Processing & Management*, **18**(1), 1-13.

O'Brien, Edward, & Myers, Jerome L. (1987). The role of causal connections in the retrieval of text. *Memory & Cognition*, **15**(5), 419-427.

Orvis, B.R., Cunningham, J.D., & Kelley, H.H. (1975). A closer examination of causal inference: The role of consensus, distinctiveness and consistency information. *Journal of Personality and Social Psychology*, **32**, 605-616.

Owens, David. (1992). *Causes and coincidences.* Cambridge: Cambridge University Press.

Pazzani, Michael, Dyer, Michael, & Flowers, Margot. (1987). Using prior learning to facilitate the learning of new causal theories. In J. McDermott (Ed.), *IJCAI 87: Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 277-279). San Mateo, CA: Distributed by Morgan Kaufman Publishers.

Peterson, Philip L. (1981). What causes effects? *Philosophical Studies*, **39**(2), 107-139.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1972). *A grammar of contemporary English.* London: Longman.

Ranganathan, S.R. (1965). *The Colon Classification*. New Brunswick, N.J.: Graduate School of Library Service, Rutgers, the State University.

Rapoport, T.R. (1990). Secondary predication and the lexical representation of verbs. *Machine Translation*, **5**(1), 31-55.

Rau, Lisa. (1987). Knowledge organization and access in a conceptual information system. *Information Processing & Management*, **23**(4), 269-283.

Rau, Lisa F., Jacobs, Paul S., & Zernik, Uri. (1989). Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, **25**(4), 419-428.

Roget's international thesaurus. (1962). 3rd ed. New York: Thomas Y. Crowell Company.

Ruble, Diane M., & Feldman, Nina S. (1976). Order of consensus, distinctiveness, and consistency information and causal attributions. *Journal of Personality and Social Psychology*, **34**(5), 930-937.

Ruge, Gerda, Schwarz, Christoph, & Warner, Amy J. (1991). Effectiveness and efficiency in natural language processing for large amounts of text. *Journal of the American Society for Information Science*, **42**(6), 450-456.

Rumelhart, David E. (1979). Some problems with the notion of literal meanings. In A. Ortony, A. (Ed.), *Metaphor and Thought* (pp. 78-90). Cambridge: Cambridge University Press.

Rutherford, William E. (1970). Some observations concerning subordinate clauses in English. *Language*, **46**, 97-115.

Salton, Gerard. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, Mass.: Addison-Wesley.

Salton, Gerard, Buckley, Chris, & Smith, Maria. (1990). On the application of syntactic methodologies in automatic text analysis. *Information Processing & Management*, **26**(1), 73-92.

Salton, Gerard, & McGill, Michael J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Salton, Gerard, Yang, C.S., & Yu, C.T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, **26**(1), 33-44.

Saussure, Ferdinand de. (1959). *Course in general linguistics*. New York: Philosophical Library. (Original work published in French in 1915).

Schamber, Linda, Eisenberg, Michael, & Michael Nilan. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management*, **26**(6), 755-776.

Schank, R.C. (1982). *Dynamic memory*. New York: Cambridge University Press.

Schank, R.C., & Abelson, R.P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, N.J.: Erlbaum.

Schleppegrell, Mary J. (1991). Paratactic because. *Journal of Pragmatics*, **16**(4), 323-337.

Schubert, Lenhart, & Hwang, Chung Hee. (1989). An episodic knowledge representation for narrative texts. In R.J. Brachman, H.J. Levesque & Reiter, R. (Eds.), *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning* (pp. 444-458). San Mateo, Calif.: Morgan Kaufmann Publishers.

Schustack, Miriam W., & Sternberg, Robert J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, **110**(1), 101-120.

Schwarz, Christoph. (1990a). Automatic syntactic analysis of free text. *Journal of the American Society for Information Science*, **41**(6), 408-417.

Schwarz, Christoph. (1990b). Content based text handling. *Information Processing & Management*, **26**(2), 219-226.

Selfridge, Mallory. (1989). Toward a natural language-based causal model acquisition system. *Applied Artificial Intelligence*, **3**(2-3), 107-128.

Selfridge, Mallory, Daniell, Jim, & Simmons, Dan. (1985). Learning causal models by understanding real-world natural language explanations. In *The Second Conference on Artificial Intelligence Applications: The Engineering of Knowledge-Based Systems* (pp. 378-383). Silver Spring, MD: IEEE Computer Society.

Sheridan, Paraic, & Smeaton, Alan F. (1992). The application of morpho-syntactic language processing to effective phrase matching. *Information Processing & Management*, **28**(3), 349-369.

Shibatani, Mayayoshi. (1976). The grammar of causative constructions: A conspectus. In M. Shibatani (Ed.), *Syntax and semantics. Vol. 6, The grammar of causative constructions* (pp. 1-40). New York: Academic Press.

Shultz, Thomas. (1982). Causal reasoning in the social and nonsocial realms. *Canadian Journal of Behavioral Science*, **14**(4), 307-322.

Simpson, Jane. (1983). Resultatives. In L. Levin, M. Rappaport, & A. Zaenen (Eds.), *Papers in lexical-functional grammar* (pp. 143-157). Bloomington, Indiana: Indiana University Linguistics Club.

Smeaton, A.F., & van Rijsbergen, C.J. (1988). Experiments on incorporating syntactic processing of user queries into a document retrieval strategy. In Y. Chiaramella (Ed.), *11th International Conference on Research & Development in Information Retrieval* (pp. 31-51). New York: ACM.

Smeaton, Alan F., O'Donnell, Ruairi, & Kelledy, Fergus. (1995). Indexing structures derived from syntax in TREC-3: System description. In D.K.

Harman (Ed.), *Overview of the Third Text REtrieval Conference (TREC-3)* (NIST Special Publication 500-225, pp. 55-67). Gaithersburg, MD: National Institute of Standards and Technology.

Somers, H.L. (1987). *Valency and case in computational linguistics*. Edinburgh : Edinburgh University Press.

Sosa, Ernest, & Tooley, Michael (Eds.). (1993). *Causation*. Oxford: Oxford University Press.

Strawson, Galen. (1989). *The secret connexion: Causation, realism, and David Hume*. Oxford: Clarendon Press.

Strzalkowski, Tomek, Carballo, Jose P., & Marinescu, Mihnea. (1995). Natural language information retrieval: TREC-3 report. In D.K. Harman (Ed.), *Overview of the Third Text REtrieval Conference (TREC-3)* (NIST Special Publication 500-225, pp. 39-53). Gaithersburg, MD: National Institute of Standards and Technology.

Swanson, Don R. (1986). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, **30**(1), 7-18.

Swanson, Don R. (1990). Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association*, **78**(1), 29-37.

Swanson, Don R. (1991). Complementary structures in disjoint science literatures. In A. Bookstein, Y. Chiaramella, G. Salton, & V.V. Raghavan (Eds.), *SIGIR '91: Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (pp. 280-289). New York: ACM Press.

Szeto, Yee-Kim. (1988). The semantics of causative and agentive verbs. *Cahiers Linguistiques d'Ottawa*, **16**, 1-51.

Tafti, M.H.A. (1992). Virtual reality: A promising tool for knowledge acquisition. In *'New Generation' Knowledge Engineering: IAKE '92: Proceedings Third Annual Symposium of the International Association of Knowledge Engineers* (pp. 471-479). Gaithersburg, MD: IAKE.

Thompson, Judith J. (1987). Verbs of action. *Synthese*, **72**(1), 103-122.

Tooley, Michael. (1987). *Causation: A realist approach*. Oxford: Clarendon Press.

Trabasso, Tom. (1989). Causal representation of narratives. *Reading Psychology*, **10**(1), 67-83.

Trabasso, Tom, Secco, Tom, & van den Broek, Paul. (1984). Causal cohesion and story coherence. In H. Mandl, N.L. Stein & T. Trabasso (Eds.), *Learning and comprehension of text* (pp. 83-111). Hillsdale, N.J.: Erlbaum.

Tversky, Amos, & Kahneman, Daniel. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49-72). Hillsdale, NJ: Lawrence Erlbaum Associates.

van den Broek, Paul. (1989). The effects of causal structure on the comprehension of narratives: Implications for education. *Reading Psychology*, **10**(1), 19-44.

Vendler, Zeno. (1984). Agency and causation. *Midwest Studies in Philosophy*, **9**, 371-384.

Venkatesh, Murali, Myaeng, Sung H., & Khoo, Christopher. (1994). Problem structuring and validation using information retrieval. Unpublished manuscript.

Webster's third new international dictionary of the English language, unabridged. Springfield, MA: Merriam-Webster.

Wells, Gary L., & Harvey, John H. (1977). Do people use consensus information in making causal attributions? *Journal of Personality and Social Psychology*, **35**(5), 279-293.

White, Peter A. (1990). Ideas about causation in philosophy and psychology. *Psychological Bulletin*, **108**(1), 3-18.

Wilensky, R.W. (1978). *Understanding goal-based stories* (Tech. Rep. No. 140). New Haven, CT: Yale University, Department of Computer Science.

Wilensky, R.W. (1983). *Planning and understanding: A computational approach to human reasoning*. Reading, MA: Addison-Wesley.

Wojcik, Richard Henry. (1973). The expression of causation in English clauses (Doctoral dissertation, Ohio State University, 1973). *Dissertation Abstracts International*, **34-08**, 5150A.

Yuan, Yan Tang, Chang, De Yan, & Suen, C.Y. (1994). Document processing for automatic knowledge acquisition. *IEEE Transactions on Knowledge and Data Engineering*, **6**(1), 3-21.

VITA

NAME OF AUTHOR: Christopher Soo-Guan Khoo

PLACE OF BIRTH: Penang, Malaysia

DATE OF BIRTH: October 18, 1958

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

Harvard University
University of Illinois at Urbana-Champaign

DEGREES AWARDED:

Bachelor of Arts in Engineering and Applied Science, 1982, Harvard University

Master of Science in Library and Information Science, 1988, University of Illinois at Urbana-Champaign

AWARDS AND HONORS

Beta Phi Mu
Syracuse University Fellowship, 1991-1994

PROFESSIONAL EXPERIENCE

Assistant Librarian, National University of Singapore, 1988-1991

Research Assistant, School of Information Studies, Syracuse University, 1992-1993