The Automatic Acquisition of Knowledge about Discourse Connectives

Ben Hutchinson

Doctor of Philosophy
Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh
2005

Abstract

This thesis considers the automatic acquisition of knowledge about discourse connectives. It focuses in particular on their semantic properties, and on the relationships that hold between them. There is a considerable body of theoretical and empirical work on discourse connectives. For example, Knott (1996) motivates a taxonomy of discourse connectives based on relationships between them, such as HYPONYMY and EXCLUSIVE, which are defined in terms of substitution tests. Such work requires either great theoretical insight or manual analysis of large quantities of data. As a result, to date no manual classification of English discourse connectives has achieved complete coverage. For example, Knott gives relationships between only about 18% of pairs obtained from a list of 350 discourse connectives.

This thesis explores the possibility of classifying discourse connectives automatically, based on their distributions in texts. This thesis demonstrates that state-of-the-art techniques in lexical acquisition can successfully be applied to acquiring information about discourse connectives.

Central to this thesis is the hypothesis that distributional similarity correlates positively with semantic similarity. Support for this hypothesis has previously been found for word classes such as nouns and verbs (Miller and Charles, 1991; Resnik and Diab, 2000, for example), but there has been little exploration of the degree to which it also holds for discourse connectives.

We investigate the hypothesis through a number of machine learning experiments. These experiments all use unsupervised learning techniques, in the sense that they do not require any manually annotated data, although they do make use of an automatic parser. First, we show that a range of semantic properties of discourse connectives, such as polarity and veridicality (whether or not the semantics of a connective involves some underlying negation, and whether the connective implies the truth of its arguments, respectively), can be acquired automatically with a high degree of accuracy. Second, we consider the tasks of predicting the similarity and substitutability of pairs of discourse connectives. To assist in this, we introduce a novel information theoretic function based on variance that, in combination with distributional similarity, is useful for learning such relationships. Third, we attempt to automatically construct taxonomies of discourse connectives capturing substitutability relationships. We introduce a probability model of taxonomies, and show that this can improve accuracy on learning substitutability relationships. Finally, we develop an algorithm for automatically constructing or extending such taxonomies which uses beam search to help find the optimal taxonomy.

Acknowledgements

I would like to thank my supervisors Alex Lascarides and Mirella Lapata, for all their advice, support and inspiration. Alex was a great source of encouragement, and was always speedy, generous and critical in providing feedback. Mirella continually helped to focus my mind on what the important questions were, and demanded standards of technical rigour that I can only hope that I have met. I am also grateful to my examiners Jon Oberlander and Simone Teufel for their useful feedback, and for a stimulating viva.

The benefits of interesting discussions with, and feedback from, other staff at Buccleuch Place are also gratefully acknowledged, and in particular I would like to thank Bonnie Webber, Henry Thompson, Caroline Sporleder, Frank Keller, Stephen Clark, Ellen Bard, Miles Osborne and Michael White. The support and encouragement of fellow students in the department was also a great boon. Outwith Edinburgh, I am especially grateful to Alistair Knott for many interesting email exchanges discussing aspects of his own PhD research, and for his encouragement of mine.

A PhD was not something I always intended to do. I was inspired and encouraged to pursue a PhD by several of my university lecturers and work colleagues in Sydney, and for this I would like to thank in particular Norman Wildberger, Peter Slezak, Chris Manning, Dominique Estival and Cécile Pereira.

I am indebted to my sources of funding: EPSRC Grant GR/R40036/01, a Barker Graduate Scholarship (University of Sydney Travelling Scholarship), and an Overseas Research Students Award from Universities UK. I received useful feedback from the anonymous reviewers and the audiences of workshops and conferences where I have presented my work, as well as the IGK summer schools in 2002 and 2004.

On a personal note, I would like to thank my friends and family for their support, in particular those who travelled halfway round the world to visit me, often with bags laden with Australian foodstuffs.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Ben Hutchinson*)

Dedicated to Rita Donley,
whose strength, intelligence, humour and compassion
is an inspiration.

Table of Contents

List of Figures

Chapter 1

Introduction

This thesis concerns the automatic acquisition of knowledge about discourse connectives, a class of words and multiword expressions that includes *but, even though, seeing as* and *because*. Discourse connectives are of interest to linguists because of their role in signalling relations in discourse. They are also of interest to computational linguists because a range of natural language processing tasks and applications require knowledge about discourse connectives. Knowledge about discourse connectives can be obtained manually, for example by linguistic introspection or through the detailed study of a corpus. However the manual acquisition of knowledge is a time-consuming process, and many English discourse connectives have received little or no study. For other languages, even less is known about their discourse connectives, and a great deal of work is required before automated discourse processing can be done. This thesis attempts to address this knowledge acquisition bottleneck by investigating how computers can acquire knowledge about discourse connectives automatically.

The approach adopted in this thesis is empirical and corpus-based. We see this as addressing an imbalance in the field of discourse, where large-scale empirical studies have been relatively few compared to the number of theoretical analyses proposed. Following Cruse (1986), we believe that both theory development and theoretically uncommitted empirical exploration can contribute to the ultimate goal of developing an explicit theory with both descriptive adequacy and explanatory power. This thesis aims to tread the tricky path of being theory-neutral, while providing empirical information that might be used to support or refute specific accounts of discourse connectives. Our approach is to gather statistics on the distributions of discourse connectives in natural language texts, and then to use these statistics to test explicit hypotheses about connectives.

This thesis addresses a number of different knowledge acquisition tasks, because automated or computer-assisted knowledge acquisition might proceed at different depths of analysis, depending on what types of knowledge the human user already possesses (or assumes to possess). If, for example, the user has no prior knowledge of connectives, then they might want to begin by constructing clusters of connectives which are in some sense similar. This information about connective similarity might then be used as the basis for proposing semantic features for each cluster. Alternatively, a computer might be used to predict pairs of connectives which can paraphrase each other, and this information could also be used to motivate semantic features for connectives (Knott, 1996). If the connectives of a language have already been the subject of detailed study, and the relevant semantic features of connectives for that language are well understood, then computers could be used to classify connectives according to those semantic features. Finally, if an incomplete lexicon of discourse connectives has already been organised into a hierarchical taxonomy, then this taxonomy might be automatically extended by inserting additional connectives in appropriate locations within the taxonomy. A computer-assisted variant of this last task would involve allowing a human to over-ride some or all of the computer's judgements, in which case it is useful to know how much confidence the system had in its judgements.

The remainder of this chapter motivates the automatic acquisition of lexical knowledge about discourse connectives, and presents the central claims of the thesis. An outline of the thesis is then provided.

1.1 Motivation for learning about discourse connectives

Texts can contain ambiguities as to how events are ordered temporally. For example, (1.1) has two interpretations: one where the slipping precedes and causes the spilling, and one where the opposite is the case.

(1.1) The cleaner slipped. He spilt a bucket of water.

Discourse connectives can be used to explicitly signal which ordering was intended. For example, *because* can be used to indicate that the spilling caused (and thus preceded) the slipping, whereas *and* would indicate that the slipping occurred first:

(1.2) The cleaner slipped **because** he spilt a bucket of water.

(1.3) The cleaner slipped **and** he spilt a bucket of water.

It follows that discourse connectives assist in the process of interpreting discourse, and as such are important for systems that parse discourse automatically (Marcu, 2000). Similarly, discourse connectives enable Natural Language Generation applications to reduce the ambiguity of the texts they produce, and even to signal semantic relations that could never be inferred by the reader otherwise. Appropriate handling of discourse connectives can also be important for text summarisation. Consider, for example, the implications of extracting the following sentence from its original context and including it in a summary.

(1.4) **Otherwise** they will invade Iran.

The new context of this sentence within the summary could radically change its truth conditions, because *otherwise* has the interpretation "if not X, then. . . ", where X is determined from the preceding sentences, and these sentences might have come from completely different parts of the original document.

In general, the automatic processing of discourse requires many types of information about discourse connectives, including not just how they affect truth conditions, but also what presuppositions and pragmatic implicatures are involved, as well as where the semantic arguments to the connective can be found. Knowledge about relationships between connectives can also be an invaluable. For example, if one wishes to construct a paraphrase for a text by replacing one connective with another (for example to make the text easier to read), it is necessary to know which discourse connective can signal the same relations (Siddharthan, 2003).

In addition to their practical utility for Natural Language Processing, it has also been proposed that discourse connectives can be a useful source of empirical data for the development of theories of discourse coherence (Knott, 1996). The argument for this can be summarised as follows. Firstly, it is likely that people actually use coherence relations when they process texts. Evidence in support of this comes from a wide range of psycholinguistic experiments showing that coherence relations affect the speed at which texts are read (e.g. Louwerse, 2001; Caron et al., 1988; Sanders and Noordman, 2000; Noordman and Vonk, 1997; Townsend, 1983). Furthermore, the fact that readers make use of coherence relations explains how they readily interpret texts such as (1.1). Secondly, if the communication of coherence is of importance to the writer, then it is likely that linguistic devices exist in order to signal those relations explicitly. The utility of signalling relations explicitly is illustrated by comparing the ambiguous (1.1) with the unambiguous (1.2) and (1.3). Thirdly, it follows that discourse connectives can be taken as evidence for the coherence relations that people use when processing texts.

Thus, the study and classification of discourse connectives is important for both practical tasks and theory development. However the manual classification of discourse connectives is a laborious task, to the extent that there has been no complete study of English connectives. In possibly the largest study of discourse connectives to date, Knott (1996) compiles a list of about 350 connectives. Of these, he analyses just forty four, or about 13%, in terms of semantic features and even for these he is "not sure of the value of" 28% of features (p. 200). Knott also produces a taxonomy illustrating whether pairs of connectives are substitutable in the given discourse contexts. This taxonomy contains about 150 connectives, however this represents only 18% of all possible pairs of connectives. There is thus still much work to be done for English discourse connectives, and much more for other languages whose discourse connectives have received far less study.

This thesis proposes that computer-assisted acquisition of information about connectives may be a solution to the knowledge acquisition bottleneck. Towards this end, it investigates the acquisition of knowledge both about individual connectives, and about relationships between connectives. The latter includes knowledge about which discourse connectives have similar meanings, knowledge of which discourse connectives can be used as paraphrases for which others, and knowledge of how the lexicon of discourse connectives can be represented in a taxonomy. At a time when statistical approaches to discourse processing are becoming more common (Marcu, 1999; Marcu and Echihabi, 2002; Sporleder and Lascarides, 2004; Lapata and Lascarides, 2004; Girju and Woods, 2005), this thesis investigates the fundamental relationship between the meaning of discourse connectives and their empirical distributions.

1.2 Contributions

The lexicon plays a crucial role in both the theory and practice of Natural Language Processing (NLP). For example, formal theories of grammar place an increasing amount of importance on the role of the lexicon (e.g. Lexical Functional Grammar, Head-driven Phrase Structure Grammar, Categorial Grammar), and the majority of state-of-the-art approaches to NLP tasks utilise lexical information to some extent, often in the form of corpus frequencies. However the size and complexity of the lexicon presents challenges for the manual development of lexical resources. The field of Automatic Lexical Acquisition aims to address this bottleneck by acquiring lexical resources automatically. Central to this enterprise has been an assumption that has been succinctly expressed as follows:

> You shall know a word by the company it keeps! (Firth, 1957, p. 11)

Or, to be long winded, the meaning of a word can be known from the words that occur near to it. Underlying Automatic Lexical Acquisition is the so-called Distributional Hypothesis, which states that if two words are semantically similar then they will also have similar empirical distributions (Harris, 1970). The converse is not guaranteed to hold, yet lexical co-occurrence-based models of distributional similarity have often been used as fairly successful predictors of semantic similarity (for example Grefenstette, 1994). The study of discourse connectives from this perspective is novel, as previous studies have focused overwhelmingly on nouns (e.g. Rubenstein and Goodenough, 1965; Miller and Charles, 1991), and to a lesser degree verbs (Resnik and Diab, 2000).

This thesis makes four main contributions. Two of these concern the empirical study of discourse connectives, while the other two are technical advances for improving the automatic acquisition of lexical knowledge.

This thesis' first contribution is the demonstration that the Distributional Hypothesis holds for discourse connectives. That is, semantically similar discourse connectives display similar patterns of lexical co-occurrences. That this should hold for discourse connectives is not obvious, given that connectives signal coherence relations that are, in general, sensitive to a wide variety of deep semantic and pragmatic aspects of the discourse context. For example, the interpretation of the discourse connective *when* is sensitive to subtle aspects of event structure (Moens and Steedman, 1988), as illustrated by its different interpretations in (1.5) and (1.6). In the former, using materials is concurrent with building the bridge, whereas in the latter the problems are solved only after the bridge is built.

(1.5) **When** they built the 29th street bridge, they used the best materials.

(1.6) **When** they built the 29th street bridge, they solved most of their traffic problems.

Discourse connectives can also be sensitive to expectations regarding causality and enablement. For example, the connective *even though* is appropriate in (1.7) only because one expects that normally if someone is part-time then they do not do more work.

(1.7) Sue does more work than the rest of us, **even though** she's part time.

These types of appropriateness conditions for discourse connectives can perhaps be felt most acutely when there is a mismatch between the expectations of the hearer and the speaker. Upon hearing (1.8), a hearer would infer that the speaker's expectations regarding causality have been violated.

(1.8) Sue does more work than the rest of us, **even though** she's Australian.

In addition, some discourse connectives require that the propositions that they relate have similar semantic structures (Kehler, 2002; Asher and Lascarides, 2003). Consider, for example, the difference in acceptability between the following:

(1.9) Bill went to the store, and Hilary **also** went shopping.

(1.10) % Bill went to the store, and Hilary **also** got upset.

Despite the sensitivity of discourse connectives to deep aspects of the context, this thesis demonstrates that their distributional similarity relates to semantic similarity. We do this both by conducting a series of experiments involving human subjects' judgements on connective similarity, and by showing that semantic features of connectives can be predicted from their lexical co-occurrences.

The second contribution concerns the methods used in the thesis. Discourse connectives have not previously been shown to be amenable to automatic methods for analysing their distributions, as is common for other parts of speech such as nouns and verbs. We show that automated corpus analysis techniques can be successfully applied to obtain useful distributional representations of discourse connectives. We demonstrate that the world wide web can be used as a reliable source of data, and that the use of a parser enables discourse connectives to be identified with high enough accuracy for useful co-occurrence distributions to be obtained. This shows that a knowledge-lean approach to analysing the distributions of discourse connectives is possible, without the requirement of an annotated corpus. This is important because although an annotated corpus of discourse connectives have been developed for English (Miltsakaki et al., 2004), this resource does not cover all connectives, and for other languages annotated corpora are even rarer.

The third contribution is a new method for comparing distributions of lexical items. A great number of techniques have been proposed for calculating the similarity of co-occurrence distributions. These typically involve comparing the posterior likelihoods of different co-occurrences, and all output a single number representing overall similarity. All the complex differences between two probability distributions are thus reduced to a single number, with a great amount of information lost in the process. This thesis introduces a new function for comparing probability distributions which is orthogonal to distributional similarity. In combination with a distributional similarity function, it thus enables one to build a two-dimensional

picture of how distributions differ. We demonstrate that this can improve accuracy on lexical acquisition tasks involving discourse connectives.

The fourth main contribution concerns the acquisition of relationships between lexical items. In general, the relationship of two lexical items to each other is constrained by their relationships to additional lexical items. These constraints can be logical, for example if two words are synonymous then they must stand in the same relationship to all other lexical items. Alternatively, they can be soft constraints, representing statistical tendencies in the lexicon. A statistical model of the lexicon is introduced that incorporates both types of constraints. We demonstrate that the model can successfully be deployed to learn relationships between discourse connectives.

1.3 Scope and terminology

This thesis will consider only English discourse connectives, and furthermore will postpone for future research the problems caused by polysemous discourse connectives. It thus does not consider connectives such as *while*, which can signal either contrast or temporal overlap. However it does include connectives such as *when* which do not fully specify the relation they signal. (Whether *when* signals temporal overlap or temporal succession is predictable from event structures (Moens and Steedman, 1988).) Indeed, identifying such cases of imprecision is the subject of experiments in Chapter 5. This distinction between the ambiguity and underspecification of discourse connectives is analogous to the same distinction for nouns, for example, and can present the same challenges. If a corpus of discourse connectives were annotated with the intended sense of each connective, then the techniques developed in this thesis could readily be applied to the individual senses.

This thesis also places restrictions on the syntactic category of the connectives it is concerned with. Specifically, it is only concerned with connectives which syntactically relate clauses. These include coordinating conjunctions (e.g. *but*) and a range of subordinators including conjunctions (e.g. *because*) as well as phrases introducing adverbial clauses (e.g. *now that, given that, for the reason that*). This thesis is not concerned with adverbial phrases which signal relations between the semantic contents of the clause they appear in and a second anaphorically determined argument (e.g. *meanwhile, as a result, moreover*) (Webber et al., 2003). We shall refer to this latter class of items as *discourse adverbials*, whereas *discourse connectives* will unambiguously be used to denote the former class. The term *discourse markers* will be used occasionally to denote the union of the two classes.

Finally, this thesis only concerns the acquisition of knowledge about discourse connective types, rather than tokens. By this, we mean that we do not acquire information about specific instances of discourse connectives in context, for example we do not attempt to disambiguate polysemous connectives, or to predict whether two discourse connectives are equivalent in a given context. Instead, our aim is to learn properties of the lexicon of discourse connectives (cf. Halliday and Hasan, 1976; Martin, 1992; Knott, 1996). However, in order to do this, we consider the contexts in which individual discourse connective tokens occur.

1.4 Overview of the thesis

This thesis has two main parts. The first part (Chapters 2 and 3) presents background material for the machine learning experiments, including theoretical and practical aspects. The second part (Chapters 4, 5 and 6) presents the experiments into acquiring knowledge about discourse connectives.

Chapter 2 provides the theoretical background to the thesis. Since much of the research on discourse connectives is couched in terms of theories of discourse coherence and rhetorical relations, we survey a range of theories of discourse relations. The diversity of competing theories provides testament to the complexity of the linguistic phenomena considered here. Some differences between theories appear merely cosmetic, while others are more substantive and arise from different motivations for developing theories of coherence. Theories also differ in the importance they place on the explicit signalling of relations. Some ignore discourse connectives completely, while others assign them a crucial role in theory development. Practical applications involving discourse connectives are then surveyed. In particular, the challenge of generating appropriate discourse connectives received a great deal of attention in the early to mid 1990s. Applying statistical techniques to parsing discourse has also been the subject of recent interest. The chapter concludes by introducing the main machine learning methods that are used in the experiments.

Chapter 3 addresses the data necessary for the experiments, describing in detail both what data was required and how it was obtained. Three main types of data are discussed. Firstly, the empirical methodology necessitates having a database of sentences illustrating the use of each connective. In order to construct this database, methods for detecting the presence of discourse connectives automatically are introduced. In order to obtain sufficient quantities of example sentences, a methodology for obtaining sentences from the web is used. Secondly, the machine learning experiments require that the context of each instance of a connective be

represented using a discrete set of features. Two main classes of features are introduced. The first consists of lexical co-occurrences, while the second consists of abstract linguistic features such as tense, mood and negation. Thirdly, for evaluation purposes we require gold standard judgements of relationships between discourse connectives. A taxonomy developed by Knott is extended manually to provide sufficient amounts of gold-standard data.

In general, automatic lexical acquisition can involve tasks into learning i) properties of individual lexical items, ii) relationships between pairs of lexical items, or iii) taxonomic structures representing relationships between multiple lexical items. Experiments involving these three task types constitute the next three chapters of the thesis.

Chapter 4 presents experiments into acquiring knowledge about individual discourse connectives. Knowledge acquisition is interpreted in terms of classification tasks, with different classes representing different semantic properties. Four experiments are carried out, based on semantic properties that are recurrent in the literature on coherence relations. The experiments concern polarity (roughly, whether or not there is some underlying contrast or a defeated expectation), veridicality (whether or not the related sentences are implied to be true), the basic type of relation being signalled (e.g. temporal or causal), and, finally, the direction of causality or temporal ordering. The experiments constitute support for the claim that automated corpus analysis techniques can be successfully applied to acquiring semantic information about discourse connectives.

Chapter 5 addresses the problem of learning pairwise relationships between discourse connectives. Two types of relationships are considered. The first is semantic similarity. Judgements on connective similarity are elicited from subjects, and are found to correlate significantly with the distributional similarity of the connectives. The second type of relationship is substitutability, i.e. the ability to create a paraphrase by using one discourse connective in place of another. Adopting a classification of substitutability due to Knott (1996), we find significant interactions between substitutability and both distributional similarity and the similarity ratings of subjects. We then consider the problem of predicting substitutability automatically. To do this, we introduce a new function for comparing empirical distributions of connectives. The new function measures the variation in differences between two probability distributions, and is found to assist in predicting substitutability. The experiments support the claim that the Distributional Hypothesis holds for discourse connectives.

Chapter 6 extends the techniques of the previous chapter to the task of learning sets of relationships between multiple discourse connectives. Since taxonomies enable the efficient and compact representation of many relationships, we present this task as one of taxonomy

extension. We introduce a statistical model of taxonomies that takes into account global aspects of their structure. This model is applied to experiments into extending an existing taxonomy automatically. The model is found to give better performance than simpler methods which do not take into account the global structure of taxonomies.

Chapter 7 concludes the thesis by summarising its contributions, and outlining directions for future work.

1.5 Published work

Some of the work presented in this thesis has already been published. This applies to Chapter 3 (Hutchinson, 2004b), Chapter 4 (Hutchinson, 2004a), Chapter 5 (Hutchinson, 2005b,c), and Chapter 6 (Hutchinson, 2005a).

Chapter 2

Background

Theories of discourse coherence aim to explain how clauses and sentences combine to form texts. A *coherent* text is not just a sequence of random sentences; instead it contains sentences that relate to each other in some way. For example, in (2.1) the second sentence explains why John broke his leg. In contrast, the two sentences of (2.2) do not seem related in any way.

(2.1) John broke his leg. He fell down some stairs.

(2.2) ? John broke his leg. I like plums. (Knott, 1996, p. 35)

Sometimes relations in texts are signalled explicitly. For example, we could paraphrase (2.1) by conjoining the two clauses using *because*, as in (2.3). In contrast, the inclusion of *because* in (2.4) seems strange because a causal relation is being signalled despite it not being clear how such a relation could exist.

(2.3) John broke his leg because he fell down some stairs.

(2.4) ? John broke his leg because I like plums.

We will refer to the relations that hold between sentences as *discourse relations* (although they have also been called *coherence relations, rhetorical relations, rhetorical predicates* and *conjunctive relations*), and the lexical items that signal these relations as *discourse markers*. Although it is discourse markers that are our object of study, in the following section we first compare a number of theories of discourse relations, since discourse connectives have often been analysed in terms of discourse coherence relations in the literature (for example Cohen, 1984; Halliday and Hasan, 1976; Martin, 1992; Knott, 1996; Knott and Sanders, 1998; Oates,

2000). In Section 2.2 we discuss discourse markers and their applications within Natural Language Processing. The machine learning of information about discourse markers is the central concern of this dissertation, and in Section 2.3 we introduce the machine learning methods that will be used.

2.1 Discourse relations

Many theories of discourse coherence have been developed within the broad framework we have just introduced. Three main points of difference between the theories can be recognised:

1. What do discourse relations relate? We loosely stated above that in coherent texts sentences are related. However, are the objects being related the sentences themselves, or alternatively their semantic interpretations, or even the speech acts corresponding to their production?

2. What discourse relations are possible? Presumably there is a finite set of discourse relations that can account for the principles and constraints on discourse coherence. What precisely are the relations that make up this set?

3. How do these discourse relations relate to each other? A theory of discourse relations may have greater explanatory power if it posits classes of relations sharing similar features. Along what dimensions should such classifications be made?

We now proceed to survey a number of theories of discourse relations, with emphasis on their solutions to the three questions above. This survey will show that there is a large amount of overlap between theories, but also that there are areas of conflict. In some cases, differences are purely terminological, while in others they are more substantial. The theories covered by this survey will also inform the experiments in Chapter 4, by motivating the choice of semantic properties that we will attempt to learn.

Discussing the various theories necessitates introducing a range of theory-specific technical terms, which are often spelt identically to common words with non-technical usages, e.g. *addition, basic, positive* and *satellite*. Fonts will be used to signal when theory-specific words are being used: SMALL CAPITALS will be used for the names of discourse relations (e.g. ADDITION), classes of discourse relations (e.g. BASIC), and properties of discourse relations (e.g. POSITIVE); **bold font** will be used for other technical terms within the various theories (e.g. **satellite**).

PARATACTIC	HYPOTACTIC	NEUTRAL
ALTERNATIVE	SUPPORTING	COLLECTION
RESPONSE	SETTING	COVARIANCE
	IDENTIFICATION	ADVERSATIVE

Figure 2.1: Overview of Grimes' relations

2.1.1 Some early accounts of relations between propositions

Grimes (1975) considers propositions to be of two sorts: **lexical propositions** have arguments which are related to their predicates via **semantic roles**; and **rhetorical propositions**, which do not. Rhetorical propositions take both lexical propositions and other rhetorical propositions as arguments, although they can at times be dominated themselves by a lexical proposition, as in:

(2.5) We just realised that either we will have to leave home before six or they will have to postpone the meeting

Here the lexical predicate *realise* dominates the clauses linked by *either... or*.

Grimes makes a distinction between PARATACTIC and HYPOTACTIC rhetorical predicates. The former treat all their arguments equally, whereas the latter makes one argument subordinate to the other. The distinction between PARATACTIC and HYPOTACTIC has correlates in various forms in many subsequent theories. A third class of NEUTRAL predicates can be either PARATACTIC or HYPOTACTIC, according to the context. For example, he postulates a COVARIANCE relation for both (2.6) and (2.7), but claims the former is PARATACTIC, the latter HYPOTACTIC.

(2.6) George eats garlic. Nancy therefore avoids him.

(2.7) George eats garlic, which is why Nancy avoids him.

The main categories of Grimes' classification are shown in Figure 2.1. The table does not show another distinction made by Grimes between SYMMETRIC and ASYMMETRIC rhetorical predicates. A predicate P is SYMMETRIC if whenever $P(a,b)$ then also $P(b,a)$, for example the ALTERNATIVE relation is SYMMETRIC, whereas the COVARIANCE relation (which includes conditional and causal sub-types) does not.

BASIC	ELABORATIVE
CONJOINING ($\wedge$)	PARAPHRASE
ALTERNATION ($\vee$)	ILLUSTRATION
IMPLICATION ($\rightarrow$)	DEIXIS
TEMPORAL	ATTRIBUTION

Figure 2.2: Overview of Longacre's relations

For Longacre (1983), clauses introduce what he calls **predications**, and he is particularly concerned with the logical relations that can hold between these. He defines a class of BASIC relations that includes the operations of the propositional calculus, namely CONJOINING ($\wedge$), ALTERNATION ($\vee$) and IMPLICATION ($\rightarrow$). Each of these operations is divided into subtypes, for example CONTRAST is a subtype of CONJOINING. Also included in the class of basic operations is a set of TEMPORAL relations, on the grounds that temporal relations are of particular importance to natural language. The basic operations are supplemented by four ELABORATIVE classes of relations, comprising PARAPHRASE, ILLUSTRATION, DEIXIS and ATTRIBUTION. Many of these relations also have FRUSTRATED counterparts, where a relation which is expected is not satisfied. For example, in *They set out for Paris but never arrived* involves FRUSTRATED SUCCESSION, because it is expected that setting out for Paris will be followed by arriving there.

2.1.2 Cohesion based accounts

Halliday and Hasan (1976) proposed that a property which they call **cohesion** is responsible for creating a coherent discourse from a sequence of sentences. Cohesion is a relation between linguistic devices in texts, and they propose five distinct subtypes: **reference, substitution, ellipsis, lexical cohesion** and **conjunction**. Of these, the first four can be thought of as using the previous linguistic context to aid the interpretation of a sentence. In contrast, the last type, **conjunction**, specifies "the way in which what is to follow is systematically connected to what has gone before" (p. 227). Halliday and Hasan describe two levels at which this connection can occur: the relation is EXTERNAL if it relates the semantic content of the sentences; it is INTERNAL if it relates the communicative process of producing those sentences. The distinction is illustrated in (2.8) and (2.9).

ADDITIVE	ADVERSATIVE	CAUSAL	TEMPORAL
COMPLEX	CONTRASTIVE	SPECIFIC	SEQUENTIAL
(furthermore)	*(but)*	*(it follows)*	*(then)*
APPOSITION	CORRECTION	CONDITIONAL	SIMULTANEOUS
(for instance)	*(rather)*	*(in that case)*	*(at the same time)*
COMPARISON	DISMISSAL	RESPECTIVE	CONCLUSIVE
(by contrast)	*(anyhow)*	*(in this respect)*	*(finally)*
			CORRELATIVE
			(first... then)

Figure 2.3: Overview of Halliday and Hasan's conjunctive relations

(2.8) First he stood up.
Next he inserted the key into the lock. (EXTERNAL)

(2.9) Firstly, he was unable to stand upright.
Next, he was incapable of inserting the key into the lock. (INTERNAL)

The conjunctive relations are classified into four major classes: ADDITIVE, ADVERSATIVE, CAUSAL and TEMPORAL. Each of these has many subclasses; the higher level distinctions are shown in Figure 2.3. Because of their interest in the linguistic devices that signal cohesion, they give examples of linguistic items that can signal each type of relation, and examples of these are also shown in Figure 2.3.

Martin's (1992) theory of relations bears many similarities to that of Halliday and Hasan. He adopts their distinction between EXTERNAL and INTERNAL, and the similarities between the major classes of relation can be seen by comparing Figure 2.4 with Figure 2.3. There are two major differences between the theories, however. Firstly, whereas Halliday and Hasan have a top-level category of ADVERSATIVE relations, Martin includes a similar range of phenomena in the subcategory CONCESSION of the CONSEQUENTIAL category. Secondly, the subcategory of COMPARISON has been promoted to a top-level category in Martin's taxonomy. Another point of difference is that Martin includes an additional orthogonal distinction between PARATACTIC and HYPOTACTIC relations. These can, though they need not be, signalled by coordinating and subordinating conjunctions, respectively.

ADDITIVE	COMPARATIVE	CONSEQUENTIAL	TEMPORAL
ADDITION *(furthermore)* ALTERNATION *(or)*	CONTRAST *(but)* SIMILARITY *(likewise)*	PURPOSE *(so that)* CONDITION *(if)* CONSEQUENCE *(so)* CONCESSION *(although)* MANNER *(thus)*	SUCCESSIVE *(then)* SIMULTANEOUS *(at the same time)*

Figure 2.4: Overview of Martin's relations

2.1.3 Knowledge based approaches

Hobbs (1985) is concerned with making precise the knowledge or beliefs that are required to successfully interpret a text. For example, he asks what knowledge is required to understand (2.10).

(2.10) John took a book from the shelf. He turned to the index.

In order to understand this text, the reader must know that (at least some) books contain indexes, and know what people typically do with books, and so on. Hobbs sees discourse relations as aiding the interpretation process, and defines his relations in terms of what the listener can infer if a relation holds between two segments S_0 and S_1. For example, the EXPLANATION relation is defined as:

(2.11) EXPLANATION: Infer that the state or event asserted by S_1 causes or could cause the state or event asserted by S_0.

The full list of relations is given in Figure 2.5. Four subclasses of relations are recognised, although only two of these are given names. The OCCASION relations help the speaker infer how two eventualities relate to each other in space and time. The EXPANSION relations enable the listener to make inferences about what predicates hold of what discourse entities. The EXPLANATION and BACKGROUND relations both help the listener place an event in context. All these relations refers to the contents of the segments, however the EVALUATION relation lets the speaker infer why an utterance was made. Thus discourse relations help listeners make inferences both about the world and about speakers' plans.

OCCASION	EXPANSION	—	—
CAUSE	PARALLEL	EXPLANATION	EVALUATION
ENABLEMENT	GENERALISATION	BACKGROUND	
	EXEMPLIFICATION		
	CONTRAST		

Figure 2.5: Hobbs' relations

RESEMBLANCE	CAUSE–EFFECT	CONTIGUITY
PARALLEL	RESULT	OCCASION
CONTRAST	EXPLANATION	
EXEMPLIFICATION	VIOLATED EXPECTATION	
GENERALISATION	DENIAL OF PREVENTER	
EXCEPTION		
ELABORATION		

Figure 2.6: Kehler's relations

In later work, Hobbs (1990) revises his relations slightly. The main change is to introduce a relation of VIOLATED EXPECTATION into the EXPANSION class. Kehler's (2002) set of discourse relations are closely related to those of Hobbs, for example he phrases relation definitions in terms of the inferences the relation enables. However Kehler re-organises the relations into three categories of RESEMBLANCE, CAUSE–EFFECT and CONTIGUITY. These three classes are borrowed from Hume (1748), who used them to classify relations between ideas.

2.1.4 Rhetorical Structure Theory

Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) is a theory of discourse in which **text spans** are connected by **rhetorical relations**. Text spans are generally taken to be either clauses or larger units composed of clauses. In order for a relation to exist between two text spans a number of **constraints** must be satisfied, and these constraints are given in the definitions of each relation. The relation definitions use a concept of **nuclearity** to distinguish the two related text spans. One text span is deemed to be more central to the writer's purpose,

SUBJECT MATTER		PRESENTATIONAL
INTERPRETATION	CONTRAST	JUSTIFY
RESTATEMENT	SEQUENCE	EVIDENCE
EVALUATION	Subclass of causal relations:	CONCESSION
ELABORATION	VOLITIONAL CAUSE	MOTIVATION
CIRCUMSTANCE	NON-VOLITIONAL CAUSE	ANTITHESIS
SOLUTIONHOOD	VOLITIONAL RESULT	BACKGROUND
CONDITION	NON-VOLITIONAL RESULT	ENABLEMENT
OTHERWISE	PURPOSE	
SUMMARY		

Figure 2.7: Mann and Thompson's relations

and this is called the **nucleus**. The less central span is called the **satellite**. (There are also two multinuclear relations for which neither span is deemed more central: SEQUENCE and CONTRAST.)

The constraints in the definition of each relation may make reference either to the propositional content of the spans or to the intentions that led to their utterance. Each relation definition also lists an effect which the writer intends to achieve. For example, the EVIDENCE relation is intended to increase the reader's belief of the content of the nucleus span.

The standard 23 RST relations are listed in Figure 2.7. Alternative groupings of the relations are possible, however a distinction is made between SUBJECT-MATTER and PRESENTATIONAL relations. SUBJECT-MATTER relations have the intended effect that the reader recognises that the relation in question holds, while PRESENTATIONAL relations aim to increase some inclination in the reader.

In RST, related text spans combine to form larger text spans which then become arguments to further relations. An adjacency constraint on text spans being related means that the internal structure of each text span forms a tree. A consequence of this is that a given pair of text spans can be related by no more than one relation. The discourse analyst is therefore forced to choose which relation is the most suitable. Moore and Pollack (1992) point out that this may force the analyst to choose between representing a SUBJECT-MATTER relation and a PRESENTATIONAL one, i.e. between representing the informational or intentional structure of a text. The text in (2.12) has two possible RST analyses: an EVIDENCE relation in which (a) is taken as evidence

for (b), or VOLITIONAL-CAUSE relation, in which (a) is taken to be the cause of (b).

(2.12) (a) George Bush supports big business.
(b) He's sure to veto House Bill 1711.

Moore and Pollack argue that NLP requires discourse models that can capture both information and intentional information, and it is not desirable to sacrifice one in favour of the other.

Many researchers have departed from RST's original set of relations, particularly when concerned with practical tasks. For example in order to apply RST to generation, RST relations have been subdivided, amalgamated, or new relations posited (Rösner and Stede, 1992; Scott and de Souza, 1990; Hovy et al., 1992). In their discourse annotation manual, Lynn and Marcu (2001) use a set of 57 relations based on RST. Such departures from the original theory were actually anticipated by Mann and Thompson:

> One might want to change or replace the definitions... such changes are to be expected and do not cross the definitional boundaries of RST. (Mann et al., 1992, p. 70)

Knott and Dale (1994) point out that positing an open-ended list of relations has several problems. Mann and Thompson's emphasis on descriptive adequacy means little attention is paid to what the relations actually model. Without constraints on relation definitions, analysts seem free to define any relation that can be used to describe a text. Knott and Dale suggest this leaves open the possibility of defining an "inform accident and mention fruit" relation to cover the text shown in (2.13).

(2.13) John broke his leg. I like plums.

2.1.5 Cognitive motivations for relations

Sanders et al. (1992) argue that theories of discourse relations should be cognitively plausible, since the relations should be psychologically real. They argue that a theory is therefore more attractive if it uses cognitively simple concepts, and they propose a set of four **cognitive primitives** for fulfilling this requirement. The four primitives can each take take one of two values, for example the **basic operation** primitive can be either CAUSAL or ADDITIVE. These values combine to specify sets of discourse relations, as illustrated in Table 2.1. If the **basic operation** is ADDITIVE then the **order** is undefined, but otherwise all combinations are possible. In other words, the primitives are **productive** in their ability to combine.

Basic Operation	**Source of Coherence**	**Order**	**Polarity**	**Example Relation**
CAUSAL	SEMANTIC	BASIC	POSITIVE	CAUSE–CONSEQUENCE
CAUSAL	SEMANTIC	BASIC	NEGATIVE	CONTRASTIVE CAUSE–CONSEQUENCE
CAUSAL	SEMANTIC	NONBASIC	POSITIVE	CONSEQUENCE–CAUSE
CAUSAL	SEMANTIC	NONBASIC	NEGATIVE	CONTRASTIVE CONSEQUENCE–CAUSE
CAUSAL	PRAGMATIC	BASIC	POSITIVE	ARGUMENT–CLAIM
CAUSAL	PRAGMATIC	BASIC	NEGATIVE	CONTRASTIVE ARGUMENT–CLAIM
CAUSAL	PRAGMATIC	NONBASIC	POSITIVE	CLAIM–ARGUMENT
CAUSAL	PRAGMATIC	NONBASIC	NEGATIVE	CONTRASTIVE CLAIM–ARGUMENT
ADDITIVE	SEMANTIC	—	POSITIVE	LIST
ADDITIVE	SEMANTIC	—	NEGATIVE	EXCEPTION
ADDITIVE	PRAGMATIC	—	POSITIVE	ENUMERATION
ADDITIVE	PRAGMATIC	—	NEGATIVE	CONCESSION

Table 2.1: Sanders et al.'s primitives and relations

Sanders et al.'s **source of coherence** primitive relates to a distinction between **semantic** and **pragmatic** connectives made by van Dijk (1979), from whom he adopts the terminology. For Sanders et al., SEMANTIC relations concern the propositional content of the discourse segments, while PRAGMATIC relations concern the locutionary acts of producing the segments.

Sanders et al.'s theory of discourse relations differs from most others in its lack of any distinct classes of temporal relations. The issue at hand here is what aspect of discourse meaning the set of discourse relations should account for, and Sanders et al. consider temporality to be a property of individual discourse segments. Sanders et al. also do not posit a distinct relation for alternation, and so this theory is perhaps in sharpest contrast with Longacre's, in which both ALTERNATION and TEMPORAL are basic relations.

Sanders et al.'s concern with cognitive plausibility is also shared by Knott (1996), and as

Feature	Values
source of coherence	SEMANTIC/PRAGMATIC
anchor	CAUSE-DRIVEN/RESULT-DRIVEN
pattern of instantiation	UNILATERAL/BILATERAL
focus of polarity	ANCHOR-BASED/COUNTERPART-BASED
polarity	POSITIVE/NEGATIVE
presuppositionality	PRESUPPOSE/NON-PRESUPPOSED
modal status	ACTUAL/HYPOTHETICAL
rule type	CAUSAL/INDUCTIVE

Table 2.2: Knott's primitive features

a result Knott adopts the idea of relations being defined through the productive combination of primitives. A further criterion of **exhaustivity** is also introduced by Knott, which in effect states that every relation must take a value for every relation. That is, Knott prefers not to leave values undefined (as with Sanders et al.'s **order** primitive for ADDITIVE relations). Knott also argues for an empirical approach, using discourse markers to motivate relations; this will be described in detail in Section 2.2. The primitive features that Knott proposes, shown in Table 2.2, share similarities with those of Sanders et al., as well as making new distinctions such as modality and presuppositionality.

If cognitive primitives underly discourse coherence relations, then we can expect these primitives to be universal, and so expect cross-lingual similarities between discourse markers. In a rare example of an inter-lingual comparison of discourse relations, Knott and Sanders (1998) show that a set of primitives can account for discourse paraphrasability data in both Dutch and English (Stede's work has also been concerned with the inter-lingual study of coherence relations, in his case comparing English and German (Stede, 1994; Grote et al., 1997)). In order to account for the data, Knott and Sanders extend Sanders et al.'s (1992) original set of primitives with an additional distinction between VOLITIONAL and NON-VOLITIONAL relations.

2.1.6 Distinguishing beliefs from speech acts

Sanders et al.'s (and van Dijk's) SEMANTIC/PRAGMATIC distinction is closely related to Hall-

iday and Hasan's (and Martin's) EXTERNAL/INTERNAL distinction. Sweetser (1990) makes a similar, but three-way, distinction between CONTENT, EPISTEMIC and SPEECH-ACT relations between utterances. Her CONTENT relations are essentially the same as others' SEMANTIC (EXTERNAL) ones, while EPISTEMIC and SPEECH-ACT subdivide the PRAGMATIC (INTERNAL) relations. The distinction between the two is illustrated by Sweetser:

> [t]here is a class of causal-conjunction uses in which the causality is that between premise and conclusion in the speaker's mind. . . , and there is another class of uses in which the causality actually involves the speech act itself.

For Sweetser, EPISTEMIC relations hold at the level of premises and conclusions about what is the case in the real world. Examples are given in (2.14)–(2.16).

(2.14) John is home, or somebody is picking up his newspapers.

(2.15) If John went to that party, he was trying to infuriate Miriam.

(2.16) A: Why don't you want to take me to basketweaving this Summer?
B: Well, Mary took basketweaving, and she joined a religious cult.

On the other hand, SPEECH-ACT relations hold between the utterances themselves, as in (2.17)–(2.19).

(2.17) Would you like to come round tonight? Or is your car still in the shop?

(2.18) How old are you, if it's not a cheeky question?

(2.19) Go to bed now! And no more backtalk!

2.1.7 Intention based accounts

One primary function of texts is to have some effects on the reader, for example to inform the reader of something, or motivate the reader to perform some task. Grosz and Sidner's (1986) theory of discourse includes a theory of **intentional structure**. Each **discourse segment** (a sentence or larger level unit) is assumed to have some **discourse segment purpose (DSP)**. These can be connected by one of two possible relations: DOMINANCE and SATISFACTION-PRECEDENCE, which are defined as follows:

- DSP1 DOMINATES DSP2 if the satisfaction of DSP2 is intended to partly satisfy DSP1.

- DSP1 SATISFACTION-PRECEDES DSP2 if DSP1 must be satisfied before DSP2 can be satisfied.

Knott (2001) is concerned with the nature of the objects that are related by discourse relations. He points out that connected imperatives had previously been analysed in terms of relations between speech acts (e.g. (2.19)). In particular, the related distinctions between EXTERNAL/INTERNAL (Halliday and Hasan, 1976; Martin, 1992) and SEMANTIC/PRAGMATIC (Sanders et al., 1992) seem unable to account for temporal relations between imperatives, as in:

(2.20) Peel the onions. Then chop them.

Here the temporal succession signalled by *Then* holds not between the speech acts, but between the actions to be performed. Knott appeals to the **intended effects** of utterances in an attempt to resolve this problem. Simplifying slightly, he states that 1) the intended effect of an imperative sentence is that the hearer performs some action, 2) the intended effect of an interrogative sentence is that the hearer answers the question, and 3) the intended effect of an indicative sentence is that the hearer believes its propositional content. Given two utterances U_1 and U_2, Knott gives two possibilities for the intended effect of their combination $U_1 + U_2$:

- If the intended effect of $U_1 + U_2$ is that the hearer believes that some relation R holds between the propositional content of U_1 and the propositional content of U_2, then the relation is SEMANTIC.
- If the intended effect of $U_1 + U_2$ is that some relation R holds between the intended effect of U_1 and the intended effect of U_2, then the relation is PRAGMATIC.

2.1.8 A dynamic semantic account

Segmented Discourse Representation Theory (SDRT) (Asher, 1993; Lascarides and Asher, 1993, 1999; Asher and Lascarides, 2003) is an extension of dynamic semantics that includes a theory of discourse relations. In SDRT, the semantic representations of sentences are combined in a recursive structure called a segmented DRS (SDRS), consisting of subordinate SDRSs and discourse relations indicating relations between them. The structural representations implied by the SDRS are not constrained to being trees, as is the case in RST for example, but can more generally take the form of a directed acyclic graph.

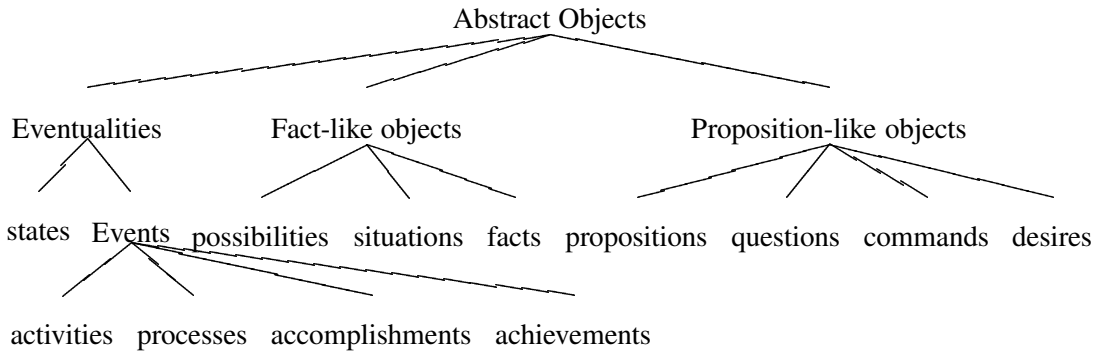


Figure 2.8: Asher's taxonomy of abstract objects

The main original motivation for SDRT was to develop a theory of discourse suitable for the analysis of abstract entity anaphora (Asher, 1993, p. 256). As has also been observed by Webber (1991), certain anaphora can refer to collections of facts and propositions already introduced in the text. For example, consider the following:

(2.21) Human life expectancy gets longer and longer. At first glance this seems like good news. But hold it. Human life is not the only thing getting longer. So are television miniseries. Well you may say, **it** only proves that Parkinson's law also fits human life: the entertainment expands to fit the time.

The referent of the boldface *it* appears to be the fact that both human life expectancy and television miniseries are both getting longer. This content is introduced by three sentences in the text: the first, the fourth and the fifth. Asher proposes that anaphora can pick out a variety of different abstract objects from the preceding discourse. His classification of abstract objects is shown in Figure 2.8.

SDRT claims that it is the **structural properties** of the text, as determined by discourse relations, that determine which such objects are available for anaphoric reference. In all, there are 38 discourse relations in the theory, of which the 13 which are applicable to monologue are shown in Figure 2.9. The relations are of two main types: content-level and text structuring. The text structuring relations require either a common or a contrasting theme, but this constraint is not required of the content-level relations. Each discourse relation is given an interpretation in the language of dynamic semantics, for example if the RESULT relation connects two SDRSs α and β then the eventuality expressed by α must cause the eventuality expressed by β. A

CONTENT-LEVEL			TEXT
VERIDICAL		NON-VERIDICAL	STRUCTURING
TOPIC	ELABORATION	DEFEASIBLE CONSEQUENCE	PARALLEL
NARRATION	EXPLANATION	CONSEQUENCE	CONTRAST
CONTINUATION	BACKGROUND	ALTERNATION	
RESULT	FBP (FOREGROUND –BACKGROUND PAIR)		

Figure 2.9: Asher and Lascarides' relations for monologue

critical distinction is made between relations that are VERIDICAL, i.e. imply the truth of the related sub-SDRS, and those that are not. Another distinction is made between COORDINATING and SUBORDINATING relations, with the former having the effect of making discourse referents less accessible to anaphora.

SDRT also includes a range of discourse relations pertaining solely to dialogue (Asher and Lascarides, 1998, 2003; Lascarides and Asher, 1999, 2004). A number of these have monologue counterparts, for example EXPLANATION$_q$ and EXPLANATION$^*_q$ connect an indicative to a question about the content of the indicative and the reason for its utterance, respectively, as illustrated by (2.22) and (2.23).

(2.22) a) A: I want to go to the party tonight.
b) B: Why? [EXPLANATION$_q$]

(2.23) a) A: It's getting late.
b) B: Aren't you enjoying yourself? [EXPLANATION$^*_q$]

EXPLANATION$^*_q$ is an example of a METATALK relation, i.e. one that holds at the speech act level. There is also a set of COGNITIVE-LEVEL relations which involve the intentions and beliefs of the interlocutors. Figure 2.10 shows the SDRT dialogue relations which do not have monologue counterparts.

This concludes our comparison of different theories of discourse relations. We have seen that many have taken up the challenge proposed by David Hume:

> Though it be too obvious to escape observation, that different ideas are connected together; I do not find that any philosopher has attempted to enumerate all the

Involving interrogatives	Involving imperatives	Divergent relations
QUESTION ANSWER PAIR (QAP)	REQUEST ELABORATION	CORRECTION
INDIRECT QAP		DISPUTE
PARTIAL QAP		COUNTEREVIDENCE
NOT ENOUGH INFORMATION		
QUESTION ELABORATION		

Figure 2.10: A selection of Asher and Lascarides' relations for dialogue

> principles of association; a subject, however, that seems to me to be worthy of curiosity. (Hume, 1748)

On the surface, there appear to be many differences between the theories. However, as has been previously pointed out (for example by Knott (1996), Louwerse (2001) and Forbes (2003)), there are also many similarities between many of the theories. Some of their differences can be viewed as differences in terminology. Other differences concern the classification of essentially the same relations, in order to draw out certain similarities, and can be considered variations in emphasis.

Our comparison of theories of discourse relations was focused on representational issues, neglecting to mention any claims about how discourse relations might be inferred by the reader. Three main approaches are mentioned in the literature, although these are not mutually exclusive. The first involves the use of various forms of non-monotonic inference (Hobbs et al., 1993; Lascarides and Asher, 1993). In Hobbs et al.'s weighted abductive approach, the discourse relation which can be proved to hold from the least costly assumptions is taken to hold. Costs are computed on the basis of numerical values which are assigned to various predicates. Although it is not clear how these numerical values should be set, Hobbs et al. suggest psychological experimentation might be used to determine relative values. SDRT uses a defeasible logic component known as DICE, and relation definitions are framed as defeasible rules. Each rule is also associated with an indefeasible axiom, for example if NARRATION holds then two events must be in a certain temporal order, and a system of deduction rules specify the interaction between the indefeasible and defeasible axioms. The second approach to recognising discourse relations involves the use of explicit linguistic signals, which we are calling discourse markers (Cohen, 1984). A third class of approaches use heuristics based on properties

of the text, such as syntactic relationships between clauses (Corston-Oliver, 1998), and how right-skewed the discourse tree is (Marcu, 1997).

2.2 Discourse markers

In this thesis we use *discourse markers* as a generic term for lexical items or multiword expressions that signal discourse relations. In this section we first discuss the grammatical categories of discourse markers in English, and then discuss their function as signals of discourse relations. Applications of discourse markers to natural language generation and discourse parsing are then reviewed.

2.2.1 Grammatical properties of English discourse markers

In English, discourse markers do not form a syntactically homogeneous group. On the contrary, they are syntactically quite varied. In this section we summarise their treatment in a large modern grammar, namely *The Cambridge Grammar of the English Language* (henceforth CGEL) (Huddleston and Pullum, 2002). Due to their syntactic diversity, CGEL does not dedicate a chapter or section to discourse markers. Instead, relevant discussion is spread throughout a large number of subsections. Because of this, page numbers will be given below to guide interested readers to the relevant sections. CGEL discusses three main syntactic classes containing discourse markers: coordinators, prepositions and connective adjuncts.

Coordinators: Prototypical coordinators are *and* and *or*. These conjoin coordinate clauses of equal status, and must appear between the coordinates (pp. 1289–1293). They can often join an unlimited number of coordinates, although this is not true of the coordinator *but* (p. 1312).

(2.24) *Kim is Irish *but* Pat is Welsh *but* Jo is Scottish.

Some even less prototypical coordinators include *so, yet* and *however* (pp. 1319–1321). While these can coordinate clauses, they can also occur in combination with other coordinators:

(2.25) There was a bus strike on, *so* we had to go by taxi.

(2.26) There was a bus strike on, *and so* we had to go by taxi.

Coordinated clauses can be of different types, for example an imperative can be coordinated with an interrogative (p. 1332):

(2.27) Come around six, *or* is that too early?

Prepositions: CGEL differs from traditional grammars in its treatment of what have traditionally been called "subordinating conjunctions". It argues that only a small subset (e.g. *that* and *whether*) of these are true subordinators, and that instead the majority, including *after, since* and *though* should be analysed as prepositions. The reason for making this distinction is that words like *after* and *since* make a clear semantic contribution, whereas the subordinator *that*, for example, does not (p. 1012). CGEL argues that the fact that words like *after* and *since* can take clausal complements is no grounds for making a primary part-of-speech distinction (pp. 1012–1013). To support this stance, they point out that some verbs take clausal complements, but that no one has suggested that these are anything other than true verbs. A distinctive feature of prepositions is that they can head non-predicative adjuncts (pp. 604–605). The distinction between predicative and non-predicative adjuncts is illustrated by (2.28) and (2.29).

(2.28) *Believing that it was a Bank Holiday*, Pat stayed at home. [predicative]

(2.29) *Assuming that the cheque bounced*, there's no money for the rent. [non-predicative]

Example (2.28) contains a predicative adjunct, as the believing is being predicated of Pat. In contrast, in (2.29) the assuming is not being predicated of anyone. So *believing* in (2.28) is a verb, whereas in (2.29) *assuming* is a preposition. Some multi-word expressions are similar to prepositions in their licensing of clausal complements, such as *on the grounds [that], for fear [that]* and *in case* (pp. 623–624). However these receive subtly different syntactic analyses:

(2.30) (PP (P On) (NP the grounds (CLAUSE that . . .)))

(2.31) (PP (P For fear) (CLAUSE that . . .))

(2.32) (PP (PP (P In) (NP case)) (CLAUSE . . .))

Finally, CGEL considers *for* and *so that* to be on the boundary between coordinators and prepositions (pp. 1321–1322). Unlike prepositions, they must occur between the clauses that they link. This distinguishes *for* from the preposition *because*:

(2.33) *Because/*For* he was exhausted, he went to bed.

Also unlike prepositional phrases, *for*-phrases and *so that*-phrases cannot be coordinated:

(2.34) *He went to bed, *for* he was exhausted, and *for* he had to get up early the next day.

(2.35) He went to bed, *because* he was exhausted, and *because* he had to get up early the next day.

However, unlike prototypical coordinators, *for* and *so that* cannot appear in multiple coordinations. Furthermore, they can only link finite clauses, for example they cannot conjoin constituents such as nouns, NPs, or VPs.

Connective adjuncts: Adverbs such as *moreover, nevertheless* and *alternatively* express a relation between the clause they occur in and the preceding text. Prototypical connective adjuncts do not impose additional truth conditions on their clause, and they cannot fall within the scope of negation, be questioned, or be focused (p. 776). One consequence of this is that connective adjuncts cannot be contrasted with each other in the way that some other types of adjuncts can:

(2.36) The sojourn did not proceed quickly, but (rather) incrementally.

(2.37) *Jill had just finished her PhD. She didn't have considerable teaching experience *moreover* but (rather) *nevertheless*.

Syntactically, connective adjuncts can be adverbial phrases or prepositional phrases, with the latter exemplified by *for this reason, by contrast, in addition, in consequence, in that case* and *as a result*. CGEL also notes that only certain connective adjuncts can be fall within negation or be the focus of a cleft clause (pp. 777-778):

(2.38) It was *for this reason/*therefore* that Ed decided to resign.

(2.39) However, Ed hadn't decided to resign *for this reason/*therefore*, but because of his disagreement with the school's policy on corporal punishment.

In addition to these three classes of discourse markers from CGEL, a further syntactic category of **phrases which take sentential complements** is distinguished by Knott (1996). These consist of matrix clauses which are missing a sentential complement, and can be of either declarative or imperative type:

(2.40) It follows that (CLAUSE . . .) [declarative]

(2.41) Suppose (CLAUSE . . .) [imperative]

In Chapter 1, we introduced the terms *discourse marker*, *discourse connective* and *discourse adverbial* that we shall use throughout this thesis. These can be considered syntactic supercategories, and their relation to the categories described above is shown in Figure 2.11. (It

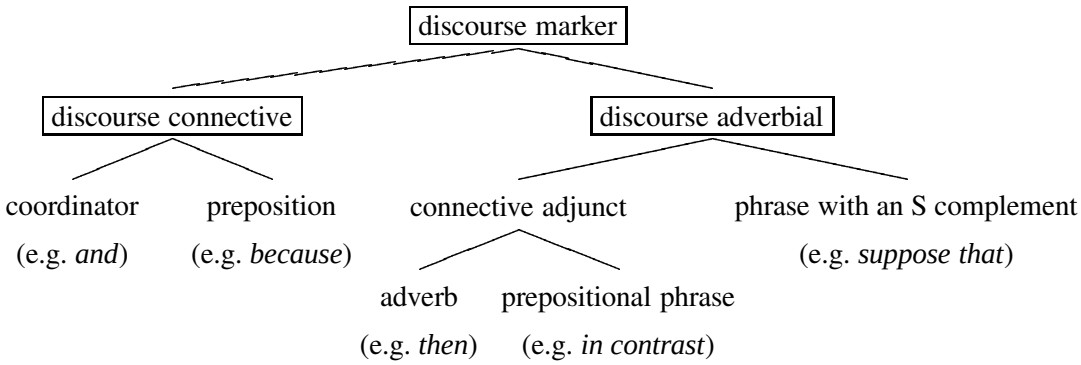


Figure 2.11: An ontology of syntactic categories of English discourse markers

should be noted that discourse markers which are phrases with S complements are not technically adverbials. However for our purposes they have a similar function to many true adverbials in that they modulate the contents of the remainder of the sentence.) The supercategories we will use (shown in boxes) are sufficiently broad to avoid the problematic borderline cases between coordinators and prepositions. When we need to, we will, in any case, refer to these basic categories as "coordinating conjunctions" and "subordinating conjunctions", respectively. This is not due to any theoretical standpoint, but merely because these terms are currently in more common use, for example they are used in *A Dictionary of Linguistics and Phonetics* (Crystal, 1997).

2.2.2 Discourse markers as signallers of discourse relations

We saw above that the many theories of discourse relations differ as to the relations they posit. They also differ in their views on the explicit signalling of those relations. It is usually accepted that an individual instance of a discourse relation need not be signalled explicitly. However Mann and Thompson (1987) go beyond this, to rule out any connection between their relations and the devices that may signal them:

> some types of rhetorical relations have no corresponding conjunctive signals (Mann and Thompson, 1987, p. 45)

Their approach may be considered "top down" in that they do not appeal to concrete linguistic signalling for the support of their theory. The mismatch between discourse markers and RST relations is illustrated by the results of a corpus annotation reported by Oates (2001). Example

tokens of each of 332 different discourse markers were selected from the BNC. If the BNC contained more than 200 instances of a discourse marker, then 200 examples were chosen at random. Each example was then annotated according to which of Mann and Thompson's original 23 RST relations it signalled. The results for the discourse markers *so* and *because* are instructive. *So* was found to signal 14 of the 23 relations, while *because* was found to signal 8. This suggests that common discourse markers may be of limited help in determining which RST relation holds.

At the other extreme, Halliday and Hasan (1976) are only interested in classifying relations that are signalled explicitly. For them, implicit signalling would not contribute any cohesion to the text, and so would lie beyond the range of their study. This text based approach can be considered "bottom up" since it takes linguistic signals as its starting point (and indeed end point too).

In between these two extremes, the majority of researchers on discourse relations have tended to accompany theoretical definitions of discourse relations with discourse markers that typically signal them (Sanders et al., 1992; Kehler, 2002, for example). Such an account does not require the assumption that the relationship between relations and markers is one-to-one however. Under one account, for example, the marker *because* is taken to signal only that the default relation is causal (Lascarides and Oberlander, 1992). Under other accounts, the many-to-one correspondence between coherence relations and discourse markers is represented through feature underspecification (Knott, 1996; Knott and Sanders, 1998). Within this middle stream, an early account of the function of discourse markers is given by Cohen (1984), who observes that discourse markers have two primary functions. Firstly, they enable readers to recognise discourse relations more quickly. Secondly, they allow the recognition of discourse relations that would not be possible in the absence of discourse markers. Note, for example, the different interpretations of the following:

(2.42) Give me your money. *Otherwise* I'll hit you.

(2.43) Give me your money. I'll hit you.

Cohen's account assumes, of course, that discourse relations are things that are recognised by the reader (whereas in RST, for example, relations are tools for the *analyst*).

Martin's (1992) theory of relations is more closely concerned with discourse markers than most within this stream. For Martin, a discourse relation is **explicit** if it is signalled by a discourse marker. **Implicit** relations are not explicitly signalled, and Martin proposes a test for such relations that uses discourse markers directly:

> As a test for the presence of an implicit connection it can be required that the connection could have been explicit. (Martin, 1992, p. 184)

For example, the acceptability of inserting *whereas* between the sentences in (2.44) is evidence that an implicit CONTRAST relation holds.

(2.44) With the big breeds of dog, they're stood on the ground, because it's easier for the judge to handle them. [*whereas*] With the smaller breeds of dog such as Corgis, all the Toy-breeds, Dachshunds and this type of thing we — as our turn comes ...

In essence, this test constitutes an important theoretical claim regarding the relationship between discourse markers and discourse relations. Martin himself admits, however, that the test does run into problems with ADDITIVE and INTERNAL relations.

In general, there have been fewer theories that make claims specifically about discourse markers (as a class) than there have been theories of discourse relations. However discourse markers have been of great importance to researchers attempting practical Natural Language Processing tasks. We therefore proceed by first surveying some of the practical applications that have used discourse markers, before considering theoretical claims that have been made about the markers themselves.

2.2.3 Applications of discourse markers in text generation

In the early 1990s, theories of coherence relations began to have a major impact on the field of natural language generation. In particular, RST and Martin's framework of relations were popular choices for implementation (Knott, 1996). These implementations were symbolic in nature, and raised practical questions concerning when relations needed to be signalled explicitly, and how to do so when they did.

Early text generation systems tended to make a one-to-one mapping between coherence relations and connectives used to signal them (McKeown, 1985, for example). However Elhadad and McKeown (1990) present a generator which chooses which discourse marker should be used to signal a relation. Their generator chooses between *but* and *although* when signalling a contrastive relation, and between *because* and *since* when signalling a causal relation. This connective selection procedure is implemented as a constraint satisfaction task, with each connective described as a set of pragmatic constraints such as "argumentative orientation" and "thematization procedure". A functional unification grammar then ensures that the constraints of the connective are consistent with features of the utterances to be connected. Grote et al.

(1997) explore related territory in a generation-oriented study of how to make (and mark) concessions in English and German. Their approach combines a model of the speaker's beliefs with a model of communicative intentions, and they present a systemic network of German discourse markers based on Martin's (1992) network for English markers.

Others have looked at signalling specific types of coherence relations. For example, Vander Linden (1994) looks at PURPOSE, RESULT and PRECONDITION relations, while Rösner and Stede (1992) and Delin et al. (1996) look at generating subject-matter relations.

Scott and de Souza (1990) propose explicit heuristics for controlling the realisation of RST relations, with the goal of making the text as easy to process as possible (given that it conveys the right message). They note that some discourse markers, such as *and*, can conjoin elements linked by a wide range of discourse relations, from which they argue that discourse markers are better thought of as just strong clues to the presence of a specific relation. On the grounds that ambiguities present processing difficulties, they advocate the use of more specific discourse markers, rather than more general ones. Furthermore, on the hypothesis that discourse relations expressed within a sentence are easier to understand than those that hold across sentences, they advocate the use of conjunctions for signalling relations in generation.

Oberlander and Lascarides (1991) make the observation that coherence relations need not be signalled using discourse markers when they can readily be inferred. Their model for discourse generation takes this inferability into account when producing candidate utterances. This is developed further by Lascarides and Oberlander (1992), who point out that texts with implicit relations are preferable to ones that use discourse markers, and define a **laconic** discourse as one which allows some inferable relations to remain implicit. They propose a method of using abduction for realising discourse coherence relations. However since the mapping between relations and markers is not one-to-one, even when a discourse marker is used the hearer may still need to infer which relation was intended. Oberlander and Knott (1995) note that the speaker sometimes has a choice as to whether to use a more general discourse marker (e.g. *after*), or a more specific one (e.g. *as soon as*), to signal a given relation. They view the choice of discourse marker within a framework of scalar Gricean implicatures. They then hypothesise that writers might use more general discourse markers in situations where the hearer can infer the specific relation, or, alternatively, when the speaker deliberately wishes to leave the coherence relation underspecified.

Moser and Moore (1995) analyse the task of discourse marker generation in terms of three separate subtasks. These are (1) whether or not to use a discourse marker to signal a relation, (2) where to place a discourse marker, and (3) which discourse marker to use. These three

aspects of marker generation interact with each other, however they have also been studied in isolation. The first two of these subtasks are the subject of machine learning experiments by Di Eugenio et al. (1997). They manually annotated a corpus with features such as informational and intentional structures and relations, as well as syntactic relations between units. These features are then used to automatically induce a range of decision trees for deciding both whether to use a discourse marker, and where it should be placed. These decision trees suggest that the speaker's purpose is important for deciding whether or not to use a marker, while syntactic relationship is most important for deciding the placement of the marker.

Grote and Stede (1998) consider the problems of discourse marker choice in generation and propose that a specialised lexicon of discourse markers can be of assistance. They propose that the lexicon should contain syntactic, semantic and pragmatic features, however the main grouping criterion of the lexicon is that of function (rather than grammatical category) on the grounds that this aspect is more important from a generation perspective. The development of the lexicon has been this subject of further research by Stede and Umbach (1998) and Berger et al. (2002).

Power et al. (2003) generate documents from an underlying structure which is related closely to RST but which does not include information about the ordering of text spans or the surface realisation of predicates. They introduce a formal notion of document structure in which discourse constituents include chapters, sections, paragraphs, and "text-sentences", and define a grammar in which each constituent is embedded within a constituent one level higher. A rule for indenting constituents allows them to generate bulleted lists, and as a consequence they can generate structures in which a bulleted list is an argument to a discourse relation, e.g.:

(2.45) Elixir is safe to use because

- it has been carefully tested
- it is approved by the FDA

The representation of discourse markers that allows them to achieve this contains four features: MEANING (i.e. rhetorical relation), SYNTAX, LOCUS (whether the marker occurs in the nucleus or the satellite constituent) and SPELLING. Constraints link the syntactic type of the discourse marker to the ordering of its arguments, as well as to the types of document constituents that can be related.

The basic assumption that discourse markers make texts easier to understand has been tested in the GIRL generation system, which produces texts for people with poor literacy skills

Structural connective	**Adverbial(s)**
although, though, whereas, but	however
or, or else	otherwise
even though	still
if, if... then	suppose... then
when	then
not only... but also	also
because, since, as	hence
and	NULL

Table 2.3: Siddharthan's paraphrasing of structural connectives by adverbials

(Williams et al., 2003; Williams, 2004). The effects on readability of Moser and Moore's three marker generation subtasks (existence, position and selection) are investigated. Reading time experiments reveal that texts that use *so* to signal a coherence relation are easier to comprehend than ones that use *therefore* to signal the same relation. The position and existence of discourse markers did not produce significant effects.

Siddharthan (2003, 2005) is also concerned with producing easily readable texts, and attempts to paraphrase complex sentences by splitting them into simpler ones. In order to ensure that the same coherence relations are present in the new text, conjunctions in the original text are replaced with discourse adverbials in the output text. For example, (2.46) is paraphrased by (2.47).

(2.46) *Though* all these politicians avow their respect for genuine cases,
it's the tritest lip service.

(2.47) All these politicians avow their respect for genuine cases.
However it's the tritest lip service.

This is achieved through the simple mapping of conjunctions onto discourse adverbials shown in Table 2.3. There is an assumption here that the conjunctions and adverbials always signal the same coherence relations. This is not always the case however, for example *when* sometimes signals temporal overlap, whereas *then* cannot. It has also been claimed that *otherwise* can only sometimes be used as a paraphrase of *or* (Knott, 1996).

The generation systems discussed above have all been symbolic in nature. However there has been recent interest using statistical approaches, both at the level of sentence generation (Langkilde and Knight, 1998; Bangalore and Rambow, 2000, for example) and at the level of document structuring (Lapata, 2003; Barzilay and Lee, 2004; Althaus et al., 2004). Creswell (2003) argues that statistical sentence generation must take into account the discourse context in which the sentence appears. In particular, Creswell demonstrates that non-canonical syntactic constructions, such as topicalisations and wh-clefts, vary in likelihood depending on the discourse context. Creswell analyses a collection of sentences beginning with the discourse connectives *and*, *but* and *so*, and finds a number of significant correlations. For example, sentences are more likely to be topicalised when they are connected by *but*, as in (2.48).

(2.48) How enforced [guardianship] ever was I don't know. *But* we would insist that the payment was made to the guardian on behalf of so and so. (Creswell, 2003, p. 180)

In contrast, the connective *so* is more likely to be used when a sentence contains a wh-cleft.

2.2.4 Applications of discourse markers in discourse parsing

The development of applications for recognising discourse relations lagged behind applications that generated discourse relations. The difficulty in recognising relations automatically lies primarily in the fact that relations need not be explicitly signalled by a discourse marker. Therefore other methods for automatically recognising relations are required. Some interpretation algorithms based on various forms of non-monotonic inference have been proposed (Hobbs et al., 1993; Lascarides and Asher, 1993, for example), however these require a detailed representation of domain knowledge, which is impractical except for very narrow domains.

Marcu (1997, 2000) describes a discourse parser that automatically produces RST-style tree structures from texts. Marcu adopts several central tenets of RST, such as that discourse structures are trees, and that discourse relations hold between adjacent spans of text. However he addresses what he sees as two shortcomings of RST. Firstly, RST lacks a formal specification for allowing one to distinguish between well-formed and ill-formed rhetorical structures. Secondly, RST lacks algorithms for producing rhetorical analyses of a given text. In a manner that has strong parallels with head-driven syntax, Marcu extends traditional RST by annotating each complex constituent with the most prominent elementary constituent in the span. He then introduces a restriction that a relation can only hold between two complex constituents if it can also hold between each of their most prominent elementary constituents.

Marcu outlines a general framework of discourse parsing, and identifies three dimensions upon which discourse parsers can vary. These are (1) the type of knowledge that the parser uses, e.g. orthography, discourse markers, semantics, (2) the type of relations that the parses identifies, and (3) the type of approach used, e.g. manually written rules, automatically derived rules, or a combination. He then proceeds to describe two implementations of parsers within this framework. The first uses knowledge of discourse markers and is based on manually written rules. The idea here is to first build discourse structures for which we have evidence from discourse markers, and then try to join these disconnected structures together. The second parser uses knowledge of discourse markers, syntax and semantics, and is based on automatically derived rules, however it only identifies relations between elementary discourse units.

Both of Marcu's parsers use knowledge about discourse markers, and this knowledge is encoded in a database with 2100 entries representing discourse marker tokens. Each entry contains fields containing information about punctuation near the marker, the position of the marker within the discourse unit, and within the sentence, the types of units related by the marker, the distance between and ordering of those units, the rhetorical relation signalled, and a range of other fields. Marcu develops regular expressions for automatically identifying discourse markers, and algorithms for identifying the elementary units of a text (e.g. clauses). He notes that these two tasks are inter-related: knowing the discourse markers in a text can help identify the elementary units.

Schilder (2000, 2002) describes a discourse parser that produces segmented discourse representation structures (Asher, 1993; Asher and Lascarides, 2003). Apart from differences in the choice of discourse representation, Schilder's parser also differs from Marcu's in that it returns partially underspecified discourse structures if it is not sure about all relations. The parser starts by assuming a completely underspecified structure for the discourse. Discourse markers are then used to introduce discourse relations into the structure, however the arguments to the relations might not be completely specified. Lastly, vector representations of each constituent are calculated, and a topicality score is then calculated for each constituent by comparing it with the title of the text. These topicality scores are taken to be indicative of relative importance in the text, and further constraints are added to the representation to reflect this. The system is implemented and evaluated using a summarisation task, on which it outperforms a range of baselines.

Forbes et al. (2003) present a discourse parsing system based on lexicalised tree adjoining grammar (TAG) (Joshi, 1987), which they call D-LTAG. The underlying assumption is that discourse level semantics has compositional aspects parallel to those at the sentence level (Webber

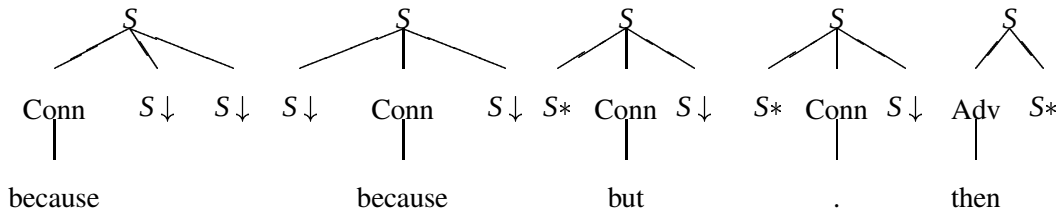


Figure 2.12: Example D-LTAG elementary trees for discourse markers

and Joshi, 1998; Webber et al., 1999, 2003). Interpreting discourse is then a combination of compositional semantics, semantic inference, and anaphora resolution. The D-LTAG parsing system produces structures which can form the input to the first of these processes, i.e. the compositional semantics. It does this by returning derivation trees, which represent the history of constructing the parse tree, and for which compositional semantics at the sentence level has already been outlined (Joshi and Vijay-Shanker, 2001; Kallmeyer and Joshi, 2003). In contrast, in the D-LTAG framework relations signalled by discourse adverbials are treated as a type of anaphora (Webber et al., 2003; Forbes, 2003). Algorithms for resolving these anaphora have not yet been implemented in the D-LTAG system.

In D-LTAG, each discourse marker is associated with an elementary tree, as illustrated in Figure 2.12. Subordinating conjunctions like *because* take two clausal arguments, as indicated by $S\downarrow$, whereas coordinating conjunctions like *but* take one clausal argument and adjoin into a previous clause. Sentence-final punctuation, such as ".", are treated in a similar manner to coordinating conjunctions, on the grounds that they connect a new sentence to the previous discourse. Finally, discourse adverbials, like *then*, adjoin to an existing *S* node but do not take sentential complements of their own. The clause the adverbial adjoins into forms one argument of the relation (or **predicate**, in their terminology), whereas the other argument needs to be resolved anaphorically.

The D-LTAG parsing system contains two stages of analysis. First, each sentence of the discourse is parsed using a sentence-level grammar and a chart-based parser (Sarkar, 2000). In the second stage discourse level parsing is performed. Clauses and discourse connectives are identified from the sentence-level parses, and a "Tree mapper" converts the elementary trees for discourse markers into the forms illustrated in Figure 2.12. This step is necessary for determining the contribution of each discourse marker to the compositional semantics.

In its current form, the D-LTAG system produces discourse representations that are primarily structural. For example, it would attempt to identify the arguments to an instance of the discourse connective like *and*, but does not disambiguate between the possible discourse relations that might be being signalled. Furthermore, if a sentence contains more than one structural connective then there can be a structural ambiguity, and it is not clear how the system selects which parse is optimal. Further structural ambiguities arise from the ability of sentence final punctuation like "." to adjoin at different nodes. Various recent research has also focused on identifying other structural relationships in discourse using machine learning methods. Thanh et al. (2004) identify the Elementary Discourse Units within a sentence using syntactic information and discourse markers. Their segmenter also predicts whether each Unit is a nucleus or a satellite of an RST relation (as does Marcu (1999)). Sporleder and Lapata (2004) use machine learning methods to automatically predict where paragraph boundaries occur in texts. So their system in effect predicts the high level discourse segmentation, and can also be useful in speech to text applications. The features they use in their machine learning experiments include the presence of discourse markers. Sporleder and Lascarides (2004) propose machine learning methods for learning high-level discourse structure. In particular, they use agglomerative clustering to produce a tree structure for a discourse in which the leaf nodes are paragraphs. Four of the machine learning features they use represent the presence of discourse markers.

Conversely, there has also been research on automatically identifying which discourse relation holds between a given pair of clauses. Marcu and Echihabi (2002) use discourse markers to collect training data for a system that identifies relations in the absence of discourse markers. Their system distinguishes between four coarse-grained relations: Contrast, Cause-Explanation, Evidence and Elaboration. A similar task is attempted by Lapata and Lascarides (2004), except they are concerned with identifying temporal relations. Their methodology is to take sentences which each contain one of eight temporal connectives, and attempt to predict from other aspects of the sentence what the connective is.

This concludes our survey of applications of discourse markers within NLP. We have seen that discourse markers have been used for a wide variety of practical applications concerned with both generating and parsing discourse. These applications all exploit the correspondence between discourse relations and discourse markers. In some cases this is done explicitly, for example when a given discourse marker is generated in order to signal a relation. In other cases the correspondence is exploited implicitly, such as when discourse markers are used as features

in a statistical discourse parser. We will now turn our attention to a theory which claims that discourse markers provide empirical evidence that can be used to motivate a theory of discourse relations.

2.2.5 Using discourse markers to motivate relations

The proliferation of theories of discourse coherence relations causes problems for developers of NLP applications as well as for discourse theorists. When developing an application, which theory of relations should one adopt? And within which framework should a theorist work to advance our knowledge of discourse? This raises meta-theoretical questions about how theories of discourse relations should be judged. Knott (1996) notes that we would like to be able to exclude theories which posit relations such as INFORM-ACCIDENT-AND-MENTION-FRUIT, even if such a relation would make (2.49) seem coherent.

(2.49) John broke his leg. I like plums.

Many researchers have proposed that discourse markers are of use in determining the set of discourse coherence relations (Ballard et al., 1971; Halliday and Hasan, 1976; Longacre, 1983; Martin, 1992). However the most developed account of how this can be done systematically has been proposed by Alistair Knott (Knott and Dale, 1994; Knott, 1996; Knott and Sanders, 1998). In Knott's account, discourse relations are held to be psychologically real, in that people actually use them when processing language. From this, Knott argues that

> ... if people actually use coherence relations when they are constructing and interpreting text, it is likely that the language they speak contains the resources to signal those relations explicitly. (Knott, 1996, p. 56)

As a result, Knott proposes that the study of discourse markers (or **cue phrases**, in his terminology), can provide evidence for the discourse coherence relations that people use. He therefore proposes an empirical methodology for studying discourse markers; in particular, his methodology focuses on relationships that hold between pairs of discourse markers. The cornerstone of Knott's methodology is a **Test for Substitutability** of discourse markers, which is summarised in Figure 2.13.

This notion of substitutability is closely related to the concept of paraphrasability. An application of the Test is illustrated by (2.50). Here *seeing as* was the original connective, however *because* can be used instead.

1. Consider any discourse marker in a text where it naturally occurs.
2. Remove the discourse marker from the text, and insert another discourse marker into the same clause as the original one (not necessarily at the same position).
3. If need be, alter the punctuation of the new discourse to make it more acceptable.
4. If need be, the new discourse can be supplemented with additional discourse markers. This might be required with pairs of markers such as *if... then* or *either... or*.
5. If it is possible to use the resulting discourse in place of the original discourse, the candidate discourse marker is **substitutable** for the original discourse marker in that context.

Figure 2.13: Knott's Test for Substitutability

(2.50) *Seeing as/because* we've got nothing but circumstantial evidence, it's going to be difficult to get a conviction. (Knott, p. 177)

This substitutability is dependent on the context however. In other contexts, for example (2.51), the substitution of *because* for *seeing as* is not valid.

(2.51) It's a fairly good piece of work, *seeing as/#because* you have been under a lot of pressure recently. (Knott, p. 177)

Similarly, there are contexts in which *because* can be used, but *seeing as* cannot be substituted for it:

(2.52) That proposal is useful, *because/#seeing as* it gives us a fallback position if the negotiations collapse. (Knott, p. 177)

Applications of the Test produce judgements of substitutability in particular contexts, and hence are concerned with individual tokens of discourse markers. Knott's next step is therefore to define relationships between discourse marker types by generalising over all contexts. This produces four possible **substitutability relationships** that can hold between two discourse markers X and Y (we follow Knott's convention of using small capitals for the relationships):

- X is a SYNONYM of Y iff X can always be substituted for Y, and vice versa.

- X is a HYPONYM of Y (and Y is a HYPERNYM of X) iff Y can always be substituted for X, but not vice versa.
- X and Y are CONTINGENTLY SUBSTITUTABLE iff each can sometimes, but not always, be substituted for the other.
- X and Y are EXCLUSIVE iff neither can ever be substituted for the other.

So, for example, (2.50)–(2.52) provide empirical evidence that *because* and *seeing as* are CONTINGENTLY SUBSTITUTABLE.

Although Knott does not point it out, his relationships are similar to Cruse's (1986) four basic "congruence relations" between lexical items: "cognitive synonymy", "hyponymy", "compatibility" and "incompatibility", respectively. (Cruse also points out that these relations are in a one-to-one correspondence with the basic relations of set theory, namely set identity, inclusion, overlap and disjunction.) The only difference is that Cruse's tool for inferring relations is semantic entailment rather than substitutability, and Knott explicitly states that the Test for Substitutability imposes stronger constraints than the preservation of truth conditions. Instead, substitutability requires the new discourse to "achieve the same goals" as the old discourse. The final stage of Knott's account is that discourse markers can be represented as sets of features. From applications of the Test for Substitutability, Knott argues for the set of features shown in Table 2.2 (repeated in the first column of Table 2.4). Features and values for some example discourse markers are shown in Table 2.4.

This framework allows Knott to explain substitutability relationships in terms of feature–values. SYNONYMY, for example, equates to having the same feature–value pairs. For example, *but* and *yet* have the same features in Table 2.4. At the other extreme, if two discourse markers are EXCLUSIVE then they take different values for the same feature. So *despite this* and *whereas* are EXCLUSIVE because they take different values for three features: **source of coherence**, **pattern of instantiation** and **rule type**. An interpretation of these differences can be obtained by consulting Knott's definitions of the features. For example, concerning the **source of coherence** feature, *despite this* signals a relation that holds between the intended effects of two utterances, while *whereas* signals a relation that holds between the propositional contents of two utterances. HYPONYMY is treated as an example of feature–value underspecification: the HYPONYM has all the feature–values of the HYPERNYM, plus others as well. This is illustrated by *but* and *despite this* in Table 2.4. Finally, if the feature–values of two discourse markers are consistent, but neither subsumes the other, the markers are CONTINGENTLY

Feature	Discourse marker		
	but/yet	*despite this*	*whereas*
source of coherence	—	PRAGMATIC	SEMANTIC
anchor of relation	—	CAUSE-DRIVEN	
pattern of instantiation	—	BILATERAL	UNILATERAL
focus of polarity	COUNTERPART-BASED	COUNTERPART-BASED	
polarity of relation	NEGATIVE	NEGATIVE	NEGATIVE
presuppositionality	NON	NON	NON
modal status	ACTUAL	ACTUAL	ACTUAL
rule type	—	CAUSAL	INDUCTIVE

Table 2.4: Knott's features for a selection of similar discourse markers. Unspecified features are denoted by "—", whereas empty cells indicate that the value is uncertain.

SUBSTITUTABLE.

Knott uses his methodology to construct a taxonomy representing substitutability relationships between 152 discourse markers. However his data-driven approach presents a number of challenges in practice. For example, the definitions of the four substitutability relationships contain generalisations over all possible contexts. As a result, three of these relationships cannot be empirically verified, only falsified. In Chapter 3 we elaborate on this and other challenges encountered when applying Knott's methodology in order to extend his taxonomy.

2.3 Machine learning methods

In this final section of the chapter, we introduce the main machine learning techniques and models that will be used in the experiments in Chapters 4 to 6. The classification techniques all rely on the automatic comparison of probability distributions representing co-occurrences of discourse connectives with various linguistic features. In the simplest cases, these features will represent occurrences of particular words in the clauses related by the connective. For example, a probability distribution p representing co-occurrences with verbs will have the following properties (Manning and Schütze, 1999):

- $p(v) \in [0,1]$ for all verbs v in $\mathcal{V}$, the set of all verbs.
- $\sum_{v \in \mathcal{V}} P(v) = 1$

We therefore proceed by first introducing the methods by which probability distributions will be compared, before considering the machine learning algorithms that use them.

2.3.1 Functions of probability distributions

There have been proposed a large number of functions for estimating the similarity (or difference) of two probability distributions p and q. For a review of a wide range of functions see Lee (1999) or Weeds (2003); here we introduce just the functions that are used in later chapters.

The relative entropy, or Kullback-Leibler (KL) divergence, measures how different two probability distributions are. Formally, it is the difference between the cross-entropy between p and q, and the entropy of p .

$$D(p||q) = -\sum_x p(x) \log q(x) - H(p) \tag{2.53}$$

$$= \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{2.54}$$

The cross-entropy of p and q takes a value which is at most the entropy of p, and it only obtains this value when $p = q$. Hence KL divergence is always non-negative, although it has no upper bound.

KL divergence is a useful function to use, because it can be interpreted informally as the expected amount of surprise when substituting one word for another. As such, a comparison can be made with Knott's Test for Substitutability, as will be discussed in detail in Chapter 5. However one potential problem in using KL divergence is that it may not be defined if there is an x such that $q(x) = 0$. In order to avoid this, Lee (1999, 2001) has defined an "skew divergence" variant which uses p to smooth the distribution of q:

$$s_\alpha(q,p) = D(p||\alpha q + (1-\alpha)p) \tag{2.55}$$

Due to the greater robustness of skew divergence, we will always use this function in preference to relative entropy. Another reason for using skew divergence is that it has been found to perform well on tasks involving lexical similarity (Lee, 2001). Since in our experiments we will *always* use the skewed variant, we will commonly simply (albeit slightly misleadingly) refer to $s_\alpha(q,p)$ as "KL divergence", since there can be no possibility of confusion. This is

intended to minimise the divergence between the reader's attention and the fact that what we are estimating relates closely to lexical substitution.

Two other functions for estimating the similarity of two probability distributions are also used. The first is the Euclidean distance function L_2, shown in (2.56), applied to probability distributions.

$$L_2(p,q) = \sqrt{\sum_x (p(x) - q(x))^2} \tag{2.56}$$

The second, $Jacc_t$, is a t-test weighted adaption of the Jaccard coefficient (Curran and Moens, 2002a). In it basic form, the Jaccard coefficient is essentially a measure of how much two distributions overlap. The t-test variant weights co-occurrences by the strength of their collocation, using the following function of two words w_i and x:

$$wt(w_i,x) = \frac{p(w_i,x) - p(w_i)p(x)}{\sqrt{p(w_i)p(x)}} \tag{2.57}$$

This is then used define the weighted version of the Jaccard coefficient, as shown in (2.58). The words associated with distributions p and q are indicated by w_p and w_q, respectively.

$$Jacc_t(p,q) = \frac{\sum_x min(wt(w_p,x), wt(w_q,x))}{\sum_x max(wt(w_p,x), wt(w_q,x))} \tag{2.58}$$

Like Kullback-Leibler divergence, $Jacc_t$ has previously been found to give good results on tasks involving lexical similarity. L_2 is included to simply to indicate what can be achieved using a somewhat naive function.

These three measures all estimate the degree to which two co-occurrence distributions differ, but they cannot specify the nature of the differences. For example, the differences between two co-occurrences distributions may be confined to a small set of co-occurrence types, or alternatively all co-occurrence probabilities may differ by a similar amount. In Chapter 5 we will introduce a new function which, in combination with measures of distributional similarity, allows a richer comparison of co-occurrence distributions.

2.3.2 Machine learning techniques

We now briefly introduce the various machine learning techniques that will be used in the experiments. The first of these, k Nearest Neighbour, does not attempt to make any generalisations from the training data. In contrast, the latter techniques construct models of the data and use Bayesian reasoning in order to make predictions.

Nearest Neighbour Classifiers

Instance based (also known as "memory based", or "example based") learning algorithms use specific instances of the training data to make predictions about test items (Aha et al., 1991). Because of this, instance-based methods are good at learning exceptions in data, and as such it has been argued that instance-based methods are highly desirable for Natural Language Processing (Varges and Mellish, 2001). Instance based methods include a wide range of classifiers, however we shall only be concerned with nearest neighbour classifiers. A precondition for these classifiers is that we can estimate the distance between two items. To do this, we will use the functions of probability distributions described above.

In the general case, a **k** Nearest Neighbour classifier determines the k training items which are closest to a given test item, and simply assigns the test item to whichever class is most represented amongst these k training items. This classifier makes no assumptions about the overall distribution of the items of each class. As such, the decision boundaries between classes can be very sensitive to individual items, particularly when k is small. An advantage of using Nearest Neighbour classifiers for lexical acquisition tasks is that they are easy to understand, and the reasons for their decisions can be explained by considering the sets of nearby neighbours. This can reveal interesting aspects of the distributions of lexical items. In practice, k can be arbitrarily large, however due to the relatively small sizes of our sets of training data, we will only use a 1 Nearest Neighbour classifier (1NN).

Nearest Neighbour classifiers do not generalise from the data. This has several consequences. Firstly, all instances in the training data must be kept in memory by the classifier, since any one could potentially be the nearest neighbour of a test instance. For large training sets, this memory requirement may become an issue. Similarly, a large number of distance calculations may be required to classify a test instance. Lastly, when a test instance is a large distance for all of the training instances, its classification may be somewhat arbitrary, as the prior likelihoods of classes is not taken into account.

Classifying using Gaussian functions

A Gaussian function, or normal function, has a symmetric "bell-shaped" curve, and is given by an equation of the following form:

$$n(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{2.59}$$

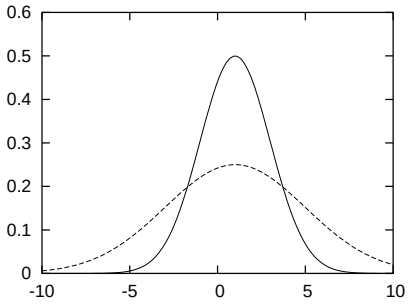


Figure 2.14: Gaussian curves with lower and higher variance

Two Gaussian functions are shown in Figure 2.14. The integral of a Gaussian functions is 1, and as a result a Gaussian function can be interpreted as a continuous probability distribution with mean μ and standard variation σ. This makes it suitable for modelling distributions of values of real-valued functions (for example distances between probability distributions). Assuming such a model, the probability of a value between x_1 and x_2 can be obtained by integrating the Gaussian function between the endpoints x_1 and x_2:

$$P([x_1,x_2]) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \tag{2.60}$$

It follows that the probability of a particular x-value, x_1 say, is always 0. However, the likelihood ratio of two x-values can be calculated by considering limits to give

$$\frac{P(x_1|\text{model } X \text{ with distribution } n(x;\mu,\sigma))}{P(x_2|\text{model } X \text{ with distribution } n(x;\mu,\sigma))} = \frac{n(x_1;\mu,\sigma)}{n(x_2;\mu,\sigma)} \tag{2.61}$$

Similarly, the ratio of the likelihoods of a given x-value being produced by two different Gaussian models with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, respectively:

$$\frac{P(x_1|\text{model } X_1 \text{ with distribution } n(x;\mu_1,\sigma_1))}{P(x_1|\text{model } X_2 \text{ with distribution } n(x;\mu_2,\sigma_2))} = \frac{n(x_1;\mu_1,\sigma_1)}{n(x_1;\mu_2,\sigma_2)} \tag{2.62}$$

Thus, Gaussian models can be used to construct a classifier by first constructing one Gaussian function per class. Let μ_i and σ_i denote the mean and standard deviation of the Gaussian function corresponding to the i'th class. To classify, we simply assign a test item with a given x-value to the class whose function assigns the maximum value:

$$Class(\text{item with } x\text{-value } x_1) = \arg\max_i n(x_1;\mu_i,\sigma_i) \tag{2.63}$$

In this framework, the probability of a given class is given by:

$$P(Class\ i|\text{item has } x\text{-value } x_1) = \frac{n(x_1;\mu_i,\sigma_i)}{\sum_j n(x_1;\mu_j,\sigma_j)} \tag{2.64}$$

Naive Bayes and ensemble methods

Given multiple classifiers for a task, perhaps based on different sets of features, there are many ways of combining their individual predictions. Naive Bayes classifiers assume that the individual classifiers are independent (i.e. knowing the given values for one does not affect our expectations regarding the values of the others). In this case, the probability of multiple events reduces to the product of the probabilities of each event, e.g.:

$$P(A,B,C,D) = P(A)P(B)P(C)P(D) \tag{2.65}$$

While Naive Bayes combines probabilities, ensemble methods combine the predictions of the individual classifiers. As such they can also be applied to classifiers which do not return probabilities, such as k nearest neighbour. An election is held, wherein each classifier votes for the class that it predicted. The class with the most votes constitutes the prediction of the ensemble.

2.4 Summary

There have been a great many theories of how discourse segments in texts are related. These theories have encouraged the development of a number of systems for either recognising or generating discourse relations. These systems have relied to a large extent on discourse markers, either as explicit signals of relations which are directly processed, or for gathering collections of examples of occurrences of discourse relations on which statistical classifiers can be trained. In both cases, a close correspondence between discourse relations and discourse markers is assumed. A range of theoretical studies have taken discourse markers as their point of departure, because discourse markers provide a concrete, empirical handle onto the study of discourse relations.

This dissertation also adopts that general research paradigm, although in our case we will be using large corpora and machine learning methods rather than introspection or manual inspection of a small number of examples. For this reason, we also introduced the main machine learning techniques that will be used. However adopting machine learning techniques imposes

three data requirements. We must have a corpus of example sentences; we must extract features from these sentences; and we must have gold standards for evaluation. The next chapter addresses these three requirements.

Chapter 3

Data requirements

In the previous chapter we reviewed previous work on discourse markers. We also introduced various machine learning methods that we will use in Chapters 4 and 5 to automatically acquire information about discourse connectives. Our main hypotheses are i) that discourse connectives with similar distributions tend to have similar properties, and ii) that discourse connectives with similar distributions are more likely to be able to paraphrase each other. However in order to obtain accurate statistical information regarding the distributions of discourse connectives, we require the following two things:

1. a large collection of texts containing discourse connectives, and
2. a set of features for representing the contexts in which discourse connectives appear.

The purpose of this chapter is to introduce both the database of texts and the set of features that we use for our experiments, and also to discuss the construction of a gold standard taxonomy for use in evaluation. The database consists of example texts for a range of different discourse connectives. Due to memory requirements, we do not store the entire text in which each connective occurs, but just enough of the text to provide the features we wish to extract. That is, the design of the database is dependent on our choice of features.

Recall from the previous chapter that discourse connectives typically signal relations between two arguments within the same sentence. More precisely, the two arguments outscope the contents of the clauses which the connective relates syntactically, for example in (3.1) the consequent of the conditional includes Edna's feeling bad and the birds getting hungry.

(3.1) If Edna forgets to fill the birdfeeder, she will feel very bad. The birds will get hungry. (Roberts, 1989)

One other case where the arguments to a relation are not contained within a sentence is when a coordinating conjunction is used sentence initially. Despite such cases, the sentence in which a discourse connective appears usually contains a lot of information about the arguments to the discourse relation that the connective signals. Since we cannot fully determine the arguments using state-of-the-art methods, we instead only extract features from the sentence in which a discourse connective appears, and we only store in our database the sentences that contain the discourse connectives, and no other context. As a consequence, we may be losing potentially important training data for learning the semantics of discourse connectives, but this is unavoidable given our inability to perform high quality discourse parsing. Our results in later chapters demonstrate that the sentence containing the connective contains enough information for a range of acquisition tasks.

The remainder of the chapter is structured as follows. In Section 3.1 we describe the construction of a database of discourse marker occurrences through extraction of sentences from both the web and the British National Corpus. This database is used as the source of statistical information in the following chapters. In Section 3.2 we introduce the range of co-occurrence features that capture aspects of the contexts in which discourse markers appear. Frequency counts of these features provide the distributional information required for the machine learning methods we described in Chapter 2. Finally, in Section 3.3 we introduce a manually constructed taxonomy that will be used to evaluate experiments on learning substitutability relationships.

3.1 A database of example sentences

The conventional starting point for empirical linguistic research is a manually constructed corpus containing texts from a variety of domains and genres, such as the British National Corpus (BNC) (Burnard, 1995). There are many advantages to using such corpora. By including a range of domains and genres, statistics on language use are unlikely to be greatly affected by any domain-specific properties of words. In addition, such corpora are often accompanied by useful annotations. These annotations may contain metalinguistic information such as the date of production and authorship of the texts, or they may contain linguistic information, such as the part of speech of each word, or specify the resolution of anaphora. Furthermore, from a scientific perspective, the widespread availability of such corpora make empirical claims regarding them easy to verify.

Nevertheless, there seems to be a growing realisation that such manually constructed and

annotated corpora are often not large enough to provide accurate statistics on many rare events. For example, despite containing texts totalling about 100 million words, the BNC contains fewer than 50 examples of most English words (Kilgarriff and Grefenstette, 2003b). Some discourse markers are among the lower frequency lexical items, for example the BNC contains just 137 occurrences of the string *seeing as*, 119 of *providing that*, 77 of *which was why*, and 29 of *for the reason that*. (And not all of these occurrences are necessarily discourse markers, as we shall discuss in Section 3.1.1.) Even for discourse markers that occur far more frequently than these ones, there is often a problem of data sparseness when trying to estimate which other words they co-occur with, since for bigrams the problem of data sparseness is further accentuated. In general, the more accurate the statistics that can be obtained the better, so various alternatives to balanced manually constructed corpora have been explored.

One approach to overcoming the data sparseness problem is to construct larger and larger corpora. This is often done by concatenating existing corpora (Curran and Moens, 2002b; Curran and Osborne, 2002; Marcu and Echihabi, 2002, for example). In so doing, care for whether or not the resulting corpus is balanced is often neglected. For example, Curran and Moens combine the balanced 100 million word BNC and the 200 million word Reuters corpus. The latter consists solely of newspaper texts, making this domain over-represented in the combined corpus.

An alternative approach is to do away with conventional corpora altogether and use other sources of linguistic data. Arguably the largest source of linguistic data is the world wide web, and the use of the web for empirical NLP has been the subject of recent interest, as is evident by a recent Special Issue of the journal Computational Linguistics (Kilgarriff and Grefenstette, 2003a). In such cases, concern for "balance" is completely discarded, indeed Kilgarriff and Grefenstette (2003b) critique the very idea of what the "representativeness" of a corpus might actually mean. In doing so, they point out that our understanding of what it means to be representative is quite primitive. They conclude: "The web is not representative of anything else. But neither are other corpora, in any well-understood sense." In fact, for certain types of linguistic data the greater size of the web seems more important than any lack of representativeness. For example, bigram statistics from the web correlate better with human plausibility judgements than bigram statistics from the balanced BNC do (Keller and Lapata, 2003).

In this thesis we adopt the second approach mentioned, i.e. we use the web as an additional source of data. We begin by discussing the problem of identifying discourse markers automatically, and propose a new algorithm for this task. We then present a method for ob-

taining sentences containing discourse markers from the web. This method incorporates the discourse marker identification algorithm. We then evaluate both the usefulness and the validity of obtaining sentences from the web. This is done by estimating the potential of the web for providing very large numbers of example sentences, and by comparing discourse marker co-occurrences in sentences obtained from the web with those in the BNC.

3.1.1 Identifying discourse markers

Many words that have uses as discourse connectives also have other uses in which they do not signal relations between clauses. This can create difficulties for NLP systems that process discourse. Some examples of discourse connectives ("DC") and homographic non-discourse connectives ("Not DC") are shown in in (3.3)–(3.15).

(3.2) Dan took off the sail cover *and* he checked the motor. [DC]

(3.3) We should swap Liz and Kim. [not DC]

(3.4) I left the party *after* Pat did. [DC]

(3.5) In the end, I didn't go to the party after all. [not DC]

(3.6) Pat likes tennis, *but* Chris likes squash. [DC]

(3.7) Pat likes tennis but not squash.[1] [not DC]

(3.8) Eat your spinach, *or* you're not getting any dessert. [DC]

(3.9) Do you want chocolate or vanilla? [not DC]

(3.10) I've been sad *since* he left. [DC]

(3.11) I haven't been home since January. [not DC]

(3.12) *Now* Pat's finally here, we can set off. [DC]

(3.13) What's the time now? [not DC]

(3.14) *Once* we had left the house, Jim began to talk more freely. [DC]

(3.15) Once upon a time there were three bears. [not DC]

[1] Here *but* is relating propositions semantically, but it is not syntactically coordinating clauses. Hence for our purposes it is not a discourse connective in this example.

The problem also persists for discourse connectives which are multiword expressions, as shown in (3.17)–(3.27).

(3.16) Pat left the party *even though* it was still early. [DC]

(3.17) Kim sanded the beam again. It was still not even though. [not DC]

(3.18) It's a fairly good piece of work, *seeing as* you have been under a lot of pressure lately. [DC]

(3.19) The problem with this conception is that it regards seeing as a purely passive activity of beholding. [not DC]

(3.20) *Assuming that* the weather holds, the picnic should be fantastic. [DC]

(3.21) I had been assuming that you would come. [not DC]

(3.22) *To the degree that* you will want to see this movie, it will be because of the surprise. [DC]

(3.23) Pat angled the cannon to the degree that I had specified. [not DC]

(3.24) You can stay up *as long as* you're quiet. [DC]

(3.25) The arm of a gorilla is nearly twice as long as its leg. [not DC]

(3.26) *Each time* the computer boots, I get a prompt. [DC]

(3.27) Will both of you each time me during my race? [not DC]

By definition, discourse connectives immediately precede clauses, since we do not count adverbials in this class. However even when a potential discourse connective is followed by a clause, there is no guarantee that it is indeed a discourse connective. For example, a word sequence such as *assuming that* can be a discourse connective, but it can also be the beginning of a predicative adjunct, as in the reading of (3.28) where it is Sue who is doing the assuming.

(3.28) Sue didn't go to work, assuming that the Bank Holiday was also a university holiday. [not DC]

Another case arises when the verb of the matrix clause takes a temporal argument, for example *took* in (3.29)–(3.32).

if preceding orthography = comma **then** *discourse*
if (part-of-speech = adverb) $\wedge$ (token = finally) **then** *discourse*
if preceding orthography = false **then** *sentential*

Figure 3.1: Example rules learned by Litman (1996) for identifying discourse markers

(3.29) John took five minutes to arrive.

(3.30) John took ages to arrive.

(3.31) John took as long as Chris to arrive.

(3.32) John took as long as Chris took to arrive.

This use of *as long as* in (3.32) differs from a true discourse connective usage of the phrase as it cannot be freely omitted:

(3.33) * John took (to arrive).

Any system that aims to process discourse markers automatically must distinguish individual tokens of discourse markers from words and phrases with identical surface forms. This task has previously been attempted for both written and spoken texts. Hirschberg and Litman (1993) proposed that discourse usages can be distinguished using prosodic features such as pitch accent and prosodic phrasing, or using textual features such as orthography and part of speech. In machine learning experiments, they used these features to induce decision trees for deciding whether a token is a discourse marker or not (Litman, 1996). Examples of rules that were learned are shown in Figure 3.1, and a decision tree trained on both prosodic and textual features achieves 84.1% accuracy. Marcu (1998, 2000) compiled a list of regular expressions for identifying discourse markers on the basis of their orthographic environments. Examples of his regular expressions are shown in Table 3.1. Marcu (1998) reports a recall of 80.8% and a precision of 89.5%.

Both these previous approaches rely on a degree of manual analysis. In the case of Hirschberg and Litman, manual annotation of tokens is required in order to learn the decision trees. In the case of Marcu, the regular expressions are developed manually after manual analysis of a corpus. In our case, we are interested in minimising reliance on manual analysis and annotation.

Discourse marker	Regular expression
Although	[␣\t\n]Although(␣|\t|\n)
because	[,][␣\t\n]+because(␣|\t|\n)
for example	[,][␣\t\n]+for[␣\t\n]example(␣|,|\t|\n)

Key: [...] defines a class of characters; (...) indicates grouping;
| indicates alternation; ␣ = space; \t = tab; \n = new line

Table 3.1: Examples of regular expressions for identifying discourse markers used by Marcu (2000). The syntax used is that for the Unix tool *lex*.

In the following section, we introduce a new procedure for identifying discourse marker tokens in a corpus. The procedure is based on general syntactic constraints that can be expected to generalise across languages.

3.1.2 A new algorithm for identifying discourse markers

In contrast to previous approaches to the problem, the new procedure that we adopt for identifying discourse marker tokens in a corpus requires no manual annotation. As a trade-off however, it does require automatic parsing of each sentence, as syntactic trees are used to rule out many non-discourse connectives. So while the development time is decreased, the processing time for each sentence is increased. However some of the co-occurrence features that we experiment with in later chapters require automatic syntactic analysis anyway, as described in Section 3.2. Thus, if parsing is required anyway, we may as well take advantage of the extra information available to us, and use it to help identify discourse markers as well. Our procedure for identifying discourse markers is summarised in Figure 3.2.

The second of these steps, correcting errors, requires explanation. The fact that any automatic parser will make errors is unavoidable. These errors will inevitably produce inaccuracies in any statistics that are based on parsed data, although hopefully this "noise" will not have important consequences. A degree of error is therefore something we have to cope with. However, manual inspection of the parse trees returned by the parser showed that there were certain parsing errors that were both common and also easily correctable. That is, it was easy for a human to see at a glance both that there was an error, and that the error could be fixed by making a simple change. An example of such a case can be seen in the fragment of a parse tree shown

1. Parse sentence using an automatic parser (Charniak, 2000).
2. Automatically correct common parsing errors.
3. Identify discourse connectives and adverbials from syntactic context:
   (a) discourse connectives precede S nodes,
   (b) discourse adverbials attach at S or VP nodes.

Figure 3.2: Algorithm for identifying discourse markers

in (3.34).

(3.34) ...(PP (IN after) (S ...)) ...

This parse fragment contains a prepositional phrase that is headed by *after* and takes a sentence as its complement. This is not a possible expansion for a PP node, so something has gone wrong. When the parser was run over the entire BNC, this parse tree fragment was found to occur 1346 times. We hypothesise that in the majority of these cases *after* is in fact a discourse connective taking a subordinate clause as a complement. We therefore make the local alteration shown in (3.35) to all these parse trees.

(3.35) ...(PP (IN after) (S ...)) ... $\longrightarrow$ ...(SBAR (IN after) (S ...)) ...

Once these changes had been made, instances of discourse markers were identified on the basis of their syntactic context. The procedure for doing this differed for discourse connectives and discourse adverbials. Since discourse connectives are always followed by a clause, connectives were identified by their proximity to S nodes in the parse tree. Figure 3.3 gives examples of patterns that were used to do this. In the case of coordinating conjunctions such as *but* and *and*, it was simply a matter of checking that the constituents they coordinated were clauses. For subordinating conjunctions (or what Huddleston and Pullum (2002) call *prepositions*), discourse connectives were identified if it was the initial word/phrase in a subordinate clause. For multi-word lexical items the situation was more complex, but the basic requirements were that there was a complete syntactic constituent (of any type) that consisted of the multi-word lexical item followed by an *S*-constituent.

```
(S ...)  (CC and) (S...)
(S ...)  (CC but) (S...)
(SBAR (IN after) (S...))
(PP (VBN given) (SBAR (IN that) (S...)))
(NP (DT the) (NN moment) (SBAR...))
(ADVP (RB as) (RB long) (SBAR (IN as) (S...)))
(PP (IN in) (SBAR (IN that) (S...)))
```

Figure 3.3: Identifying structural connectives from parse trees

As discussed above, there are phrases which are both i) homographic with discourse connectives, and ii) are followed by subordinate clauses, yet are not discourse connective (e.g. in (3.28) and (3.32)). Our algorithm therefore handles such cases incorrectly, introducing noise into our feature counts.

Discourse adverbials were identified slightly differently from discourse connectives, since they can occur at any position within a clause. A word or phrase with the surface form of a discourse adverbial was identified as one if both a) it was a complete syntactic constituent in its own right, and b) this constituent was located directly beneath either an S or a VP node. The first of these conditions rules out cases such as (3.36), while the second rules out cases such as (3.37).

(3.36) I packed my shoes (PP *in* (NP *that case* on the bed)).
⇒ *in that case* is not a discourse adverbial here (not complete constituent)

(3.37) It was estimated from detailed studies that the first landing might be possible (ADVP six months (RBR earlier)).
⇒ *earlier* is not a discourse adverbial here (not directly beneath S or VP)

3.1.3 Evaluation of the algorithm

In order to evaluate our algorithm for identifying discourse markers, we used a set of 500 sentences from the BNC which our algorithm had identified as containing discourse connectives. By evaluating only on sentences which our system identified as containing discourse connectives, we compare just the number of true positives with the number of false positives, i.e. we

Connective	Marked correct by: Judge 1	Judge 2	Inter-judge agreement Percentage	κ
after	83.0%	89.0%	88.0%	0.505
and	89.0%	89.0%	98.0%	0.898
as long as	89.0%	96.0%	91.0%	0.363
assuming that	81.0%	79.0%	92.0%	0.750
every time	91.0%	93.0%	96.0%	0.729
All 5 connectives	86.6%	89.2%	93.0%	0.671

Table 3.2: Accuracy of Sentence Analysis

calculate the precision. We do this because the web gives us a practically unbounded supply of data: it contains far more text than we could ever hope to process in practise, making high recall less important. In any case, evaluating the recall of the algorithm is not possible in the absence of gold standard annotations of discourse connectives in the BNC.

The evaluation used a set of five discourse connectives of different syntactic types: *and* (coordinating conjunction), *after* (subordinating conjunction), *as long as*, *assuming that* (V-ing+complementizer) and *every time* (quantifier+noun). By using this range of connectives, we reduced any bias towards particular syntactic constructions. For each of these connectives, 100 sentences identified as containing that connective were selected at random from the BNC. These were inspected by two human judges, who were asked whether each sentence contained the supposed discourse connective, or not, and the results are shown in Table 3.2. The judges were both computational linguists researching discourse processing: Judge 1 was the current author, Judge 2 a postdoctoral researcher. Before performing the task, the judges were given an explicit set of instructions, stating that for our purposes discourse markers relate clauses, are not subcategorised for, and do not contain verbs that are externally controlled (cf (3.28) and (3.32)). The results show that the sentence analysis module achieved accuracies of 86.6% and 89.2% with each of the two judges.[2] This is comparable to the results achieved in the previous work described above, however it does not rely on extensive manual analysis or annotation, and is likely to generalise better to other languages.

The inter-judge agreement of 93% gives an idea of the upper bound for the task. Agreement

[2]The fact that, of the two judges, the current author (Judge 1) judged the system as performing worse should allay any suspicions of experimenter bias.

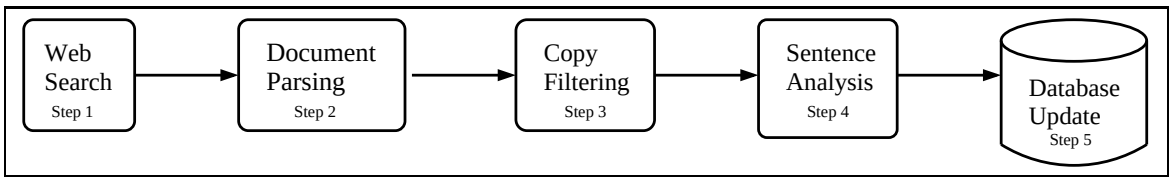


Figure 3.4: Methodology for mining the web

was also evaluated using the κ statistic (Carletta, 1996), which has the advantage of taking into account how much agreement is expected by chance. It is is defined as

$$\kappa = \frac{P_A - P_0}{1 - P_0} \tag{3.38}$$

where P_A is the probability that the judges agree in practice, and P_0 is the probability that they would have agreed by chance. Across all five discourse connectives, the judges achieved $\kappa = 0.671 (N = 500, k = 2)$, indicating a substantial level of agreement (at least by the interpretation of kappa proposed by Landis and Koch (1977); other interpretations of kappa are possible, e.g. that of Krippendorf (2004)), however κ was low for *as long as*. For this connective, a recurring source of disagreement between judges was the question of whether or not *as long as* introduced a temporal phrase that was subcategorised for by the matrix verb (cf. example (3.32)). In any case, despite this low κ score, both judges rated the accuracy in identifying *as long as* as at least 89%. The discourse connective rated as having the worst accuracy was *assuming that*, probably due to its alternative use in predicative adjuncts (as in example (3.28)).

3.1.4 A methodology for mining the web for discourse markers

We now describe a methodology for mining the web for sentences containing discourse markers. As illustrated above by examples (3.3)–(3.32), this task is made difficult by the fact that many words and expressions are ambiguous between signalling discourse relations and other uses. A necessary step is therefore the disambiguation of discourse connectives from other uses of the same words or phrases, for which we use the identification algorithm discussed above. Our methodology for mining example sentences from the web is shown schematically in Figure 3.4, and each of the main stages is summarised below. After this, we describe the implementation and evaluation of the methodology.

Step 1: Searching the web

First, a web search engine is used to find pages that may contain discourse markers, by searching for their surface forms. For example, to collect examples of the discourse connective *and*, we begin by doing a web search for `"and"`.

One complication is that many search engines restrict how many web pages may be accessed per search. For example the AltaVista search engine (`http://www.altavista.com`) only returns the top 1,000 hits. Considering that our discourse marker identification algorithm only identifies about 3% of tokens of *or* as discourse connectives, this can put a severe constraint on the number of discourse marker tokens accessible from the web. Our approach to overcoming this is to use digits as additional search terms. For example, using AltaVista we can retrieve 1,000 pages containing both *and* and the digit 1 by searching for `"and" AND 1`. Similarly, we can retrieve 1,000 more pages containing *and* (but not 1) by searching for `"and" AND NOT 1`.[3] Thus we retrieve a total of 2,000 distinct pages containing *and*. To retrieve 4,000 web pages containing *and* we make the following four searches:

- `"and" AND NOT 1 AND NOT 2`,
- `"and" AND NOT 1 AND 2`,
- `"and" AND 1 AND NOT 2`, and
- `"and" AND 1 AND 2`.

This method obviously generalises, allowing an unlimited number of web pages to be retrieved. Note that this method does make an implicit assumption that the distributions of discourse connectives and digits are independent, but this is unlikely to be harmful.

Step 2: Document parsing

The URLs returned by the search engine are downloaded and analysed automatically. An HTML parser is used to extract textual elements from the document, and punctuation heuristics are used to segment the text into sentences. Sentences not containing strings matching the relevant surface forms are filtered out.

At the end of this stage we will have a list of sentences containing both discourse and non-discourse uses of *and*, for example.

[3] The AltaVista operator `NOT` excludes documents containing the following word or phrase.

Step 3: Copy filtering

Multiple copies of identical sentences found on the web are discarded. The motivation for this is twofold. Firstly, we do not want to waste processing time by analysing the same sentence repeatedly. Secondly, we aim to avoid repetitions of a single utterance affecting our statistics. Such repetitions may occur through the mirroring of websites, syndication of news items or columns, plagiarism, or quotation (this problem is also mentioned by Kilgarriff and Grefenstette (2003a), and discussed at more length by Wilks (2004)). For example, the discourse connective *and* occurs in (3.39).

(3.39) All programmers are playwrights *and* all computers are lousy actors.

This sentence scores 1,150 hits on Google, and all these hits probably stem from a single creative utterance. We want our representation of the distribution of *and* to not be influenced by repetitions such as these.

A negative consequence of this decision is that we do not capture the frequency with which the same sentence may be created independently. For example, we lose the information that *Come and get it!* is a common use of the discourse marker *and*. This is not necessarily always a problem however, as commonly uttered sentences such as this are often idiomatic fixed expressions.

Step 4: Sentence analysis

This stage determines whether or not a sentence contains a discourse marker. A parser is run on each sentence, and the resulting parse tree is automatically analysed to determine if the previously identified surface forms are actually discourse markers, using the procedure outlined above in Figure 3.2. Sentences not containing discourse markers are discarded at this stage.

Because the web has the opportunity to provide a huge amount of training examples, we can afford to be conservative in our identification of discourse markers. However we must also be careful that in being conservative we do not collect an "unrepresentative" sample of data. In Section 3.1.6 we will evaluate the representativeness of our methodology by comparing discourse marker bigrams from the BNC with those from the web.

Step 5: Database update

Sentences identified as containing discourse markers are saved to a database, indexed by the discourse markers they contain, for later analysis, either manual or automatic. This indexing

makes it easy to use the database as a resource for analysing the distributions of particular discourse markers. The URL of the document containing the sentence is not stored, however the list of URLs output by Step 1 are saved, which enables the source of each sentence to be retrieved by re-analysis of the documents (assuming that the contents of the web pages do not change; if they do, knowing the original URL from which example sentences were collected is not useful anyway).

3.1.5 Implementation

The methodology for obtaining example sentences from the web, described above, was implemented and used to obtain data for the experiments described in following chapters. The following design decisions were made in the implementation:

- The AltaVista search engine was chosen for doing the web searching. The pragmatic reason for this was that some other search engines, such as Google (`http://www.google.com`), require prior registration for doing automated searches, and limit the number of automated searches to 1,000 per day.

- For document parsing, the HTML::Parser module from the Comprehensive Perl Archive Network (`http://www.cpan.org`) was used. This module was easy to integrate with the scripts for searching the web, also written in Perl. The heuristics for segmenting text into sentences relied on punctuation and upper/lower-case cues, as discussed by Manning and Schütze (1999).

- For parsing sentences, we use Charniak's (2000) statistical parser inspired by the principle of maximum entropy. This top-down parser is based on a probabilistic generative model, is trained on the Penn Wall Street Journal treebank (Marcus et al., 1993), and achieves 89.5% average precision/recall of labelled brackets on sentences of length ≤ 100 on the test section of that treebank.

The system was used to mine the web for sentences containing each of a set of 116 discourse connectives (taken from (Knott, 1996)), by analysing 8,000 pages containing each discourse marker's surface form. (In order to retrieve 8,000 search hits, we incorporated into the search terms all boolean combinations of the digits `1`, `2` and `3`.) These 116 include all discourse connectives used in the experiments, plus some ambiguous or unclassified ones that were not used.

Discourse marker	Tokens obtained from BNC	Tokens obtained from web	Page hits using AlltheWeb
or else	98	843	85,800,000
even if	51	13,976	230,000,000
now	23	4,361	490,000,000
for the reason that	15	1,724	75,100,000
insofar as	2	4,087	1,050,000

Table 3.3: Number of sentences identified as containing discourse markers

3.1.6 Evaluation of the web mining methodology

In this section we evaluate the methodology for mining example sentences from the web that we introduced above. In Section 3.1.3 the accuracy of the discourse marker identification stage was evaluated, so here we focus our evaluation on two additional considerations: 1) the ability of the web to provide large numbers of example sentences, and 2) the representativeness of data obtained from the web. We now describe each of these evaluations in detail.

Quantity of sentences obtainable from web

To estimate the usefulness of the web for providing large numbers of example sentences, we used five discourse connectives that were identified with a frequency of less than 100 in the BNC. These were: *or else*, *even if*, *now*, *for the reason that* and *insofar as*. For each of these, Table 3.3 lists the number of tokens that were obtained from the web using the procedure described above. Recall that this involved automatic analysis of 8,000 web pages containing string matches for a range of discourse connectives. For comparison, the table also lists how many tokens were obtained from the BNC database using the same discourse marker identification procedure. The range in quantities of tokens obtained from the web is due to a number of factors, including the likelihood of the given string being a discourse connective (for example *or else* often conjoins non-S constituents), as well as the likelihood of a connective being repeated within a document. For example, the high number of *even if* tokens obtained from the web indicates that if a web document contains one of these tokens then it probably contains more than one of them. Note that for common connectives, such as *and* and *after*, the BNC contained large numbers of tokens (371,430 and 30,551, respectively) which outnumbered the

quantities retrieved from 8,000 web pages. However for the less common connectives, the ability to search for pages matching the surface form of a connective meant that more tokens were obtained from the web than from the BNC.

Table 3.3 also lists the number of page hits for the surface form of each connective, using the AlltheWeb search engine in September 2004. This search engine was chosen as it indexes a greater proportion of the web than AltaVista does. This shows that we have analysed only a tiny fraction of the amount of data available on the web. Even for uncommon discourse markers, the web can provide hundreds of thousands of tokens, given enough processing time.

Quality of sentences mined from web

As discussed above, the web may be claimed to be unrepresentative of language use in general. However the notion of representativeness is somewhat imprecise. Here we attempted to quantify representativeness by comparing discourse marker co-occurrences obtained from the web with those in the BNC. By "discourse marker co-occurrences", we mean occurrences of more than one discourse marker within a sentence. For example, in (3.40) the connective *but*, signalling unexpectedness co-occurs with the discourse adverbial *then*, signalling temporal succession.[4]

(3.40) In one form of the myth he was killed and dismembered, *but then* his head floated. (BNC)

Co-occurrences of discourse markers such as this can indicate co-occurrences of discourse relations, and so be indicative of the discourse contexts in which discourse markers appear, as shall be discussed more in Section 3.2.2. A high correlation between such co-occurrences in the BNC and the web would indicate that sentences from the BNC and the web are at least similar in one respect. A low correlation would show that the data from the web is dissimilar to that in the BNC. Co-occurrences of discourse markers are more useful for estimating this correlation than are simple unigram frequencies of discourse markers. This is because our method of sampling from the web via searching for surface forms provides a biased sample of the web, and so unigram frequencies are not comparable with those in the BNC. By comparing bigrams, we factor out these biased unigram frequencies.

This stage of the evaluation used a different collection of discourse markers from the previous evaluations. In order to guarantee reliable statistics from the BNC, this time only five

[4]There is also a multiword connective *but then*, exemplified by: *Pat wants to come too,* but then *why wouldn't she?* However this is not the usage in (3.40).

	Correlation	#BNC bigrams	#Web bigrams
after	0.8028**	1,504	258
and	0.9259**	126,714	11,153
before	0.8555**	4,329	398
but	0.9578**	87,193	7,159
or	0.8898**	2,677	454

Table 3.4: Correlation of discourse marker bigrams (** $p < 0.001$)

high frequency discourse markers are used: *after, and, before, but* and *or*. Table 3.4 shows the correlation of discourse marker co-occurrences for each of these connectives, as well as the raw numbers of co-occurrences used in the calculations. The reason there are more BNC bigrams than Web bigrams, is that because these five discourse markers are common, they occur with a higher frequency in the BNC than they do in 8,000 web pages containing their surface strings (which may or may not be discourse markers). However, as was shown above, this inequality does not hold in general. Correlation was measured using Pearson's r, and the results indicate a high to very high degree of correlation, and are highly significant. This suggests sentences from the web contain discourse markers which are representative of their discourse contexts.

Keller and Lapata (2003) performed a similar comparison of different types of bigrams from the BNC and the web. These correlation statistics are higher on average than those found by Keller and Lapata for Adjective-Noun ($r = 0.847$), Noun-Noun ($r = 0.720$) and Verb-Object ($r = 0.762$) bigrams. This may be in part due to there being fewer distinct discourse markers than there are verbs, nouns or adjectives. However, imperfect correlation with the BNC should not be taken as a sign of imperfection, since Keller and Lapata also find that bigram statistics from the web correlate better with human plausibility judgements than do bigram statistics from the BNC. So although correlation with the BNC is an imperfect measure of representativeness, we have demonstrated that data from the web and the BNC are indeed highly correlated, on at least one dimension.

To sum up, large numbers of common discourse connectives can be obtained from conventional balanced corpora such as the BNC. For uncommon discourse connectives, the web can be mined for large numbers of tokens, and the web data correlates highly with data from the BNC, at least on one measure. For our experiments in the following chapters, we maximise our

Number of connectives (types)	140
Number of connectives (tokens)	4,588,000
Mean tokens per type	32,770
Median tokens per type	4,948
Estimated false positives	10.8–13.4%

Table 3.5: Statistics on the database of discourse connectives

data by combining the data obtained from the BNC and from the web. So for common connectives such as *and* and *but*, the BNC data has the greater effect on our statistics representing the distributions of connectives. Whereas for uncommon connectives, the data from the web has the greater contribution. Some statistics concerning the database of discourse connectives are shown in Table 3.5.

3.2 Features for machine learning experiments

At the beginning of the chapter, we presented two main requirements for the automatic statistical analysis of discourse connectives. The first of these was a collection of texts containing many discourse marker tokens. The second requirement was a selection of linguistic features for representing the context of a discourse connective. It is this requirement that we now turn our attention to.

In general, the context in which a discourse marker appears includes a large variety of factors, both linguistic and non-linguistic. For example, in (3.41) the words *that, you* and *me* all refer to items outside the text, i.e. they are deictic.

(3.41) I can't reach that bowl, *so* can you pass it to me?

Another challenge is that determining the conditions under which the use of a discourse connective is appropriate can require complex world knowledge and reasoning (Hobbs et al., 1993; Lascarides and Asher, 1993; Asher and Lascarides, 2003). The order of causation expressed by (3.42) seems appropriate, whereas that expressed by (3.43) is marked.

(3.42) The patient took an aspirin *because* he was sick.

(3.43) ? The patient was sick *because* he took an aspirin.

However if the patient is known to be allergic to aspirin then (3.43) seem perfectly acceptable, as does:

(3.44) The hyperallergic patient was sick because he took an aspirin.

Indeed, confronted with the utterance (3.43), a reader is likely to accommodate some proposition to the effect that taking aspirin can cause this patient to be ill.

Given that the appropriateness conditions for discourse markers can involve complex world knowledge, accommodation, and reasoning, it is an interesting question as to whether there are shallow linguistic features which correlate positively with occurrences of discourse markers. Such correlations would demonstrate a statistical relationship between shallow linguistic structures and the arbitrarily complex situations in which utterances are produced. Our primary hypothesis in this thesis is that such a correlation exists. Evidence from empirical studies of logical metonymy provide hope that this might be so. In verbal logical metonymy (e.g. *finished her beer*) the verb requires an event type argument (Pustejovsky, 1995). In general, logic-based methods for determining what this event is (e.g. what event it was that finished) require extensive world knowledge and reasoning. Nevertheless, corpus-based methods have been used to resolve such metonymy automatically (Lapata and Lascarides, 2002).

In order to demonstrate a correlation between the semantics of discourse connectives and their empirical distributions, we require methods for representing these representations. Our aim in this section is therefore to introduce a set of features which might be useful as shallow representations of discourse context. Our assumption is that the data will provide fewer connectives with marked arguments, such as (3.43), than connectives with more natural arguments, such as (3.42).

Since we consider discourse connectives to be signalling relations between abstract objects such as events and propositions, we can distinguish two types of contextual information: internal and external. The first consists of information about the elements that are related by the connective. It is relatively easy to obtain information about these, since syntax determines the arguments to discourse connectives (or at least part of the arguments, as discussed above). So by examining the clauses that a connective relates, we can extract a large amount of information about the events or propositions that are related. The second type of information concerns the location of the discourse relation within the overall text. This type of information is much harder to quantify, and much harder to obtain automatically. Because of this, most of our features will concern the contents of the clauses linked by a connective. However some features which we extract from within the clauses will capture aspects of the discourse context. For

example, one feature we will use is the presence of other discourse markers within the related clauses, and these can indicate the presence of other discourse-level relations in the text.

We proceed by discussing three main classes of features in turn. The first, and simplest, is word co-occurrences within the related clauses, which loosely approximate the content, or "aboutness" of the clauses. The second type of feature is the occurrence of other discourse markers, and these approximate the presence of other discourse relations within the text, as well as the double-marking of some discourse relations. The third class contains a wide variety of syntactic and semantic features that require syntactic analysis for their detection. These features are more abstract, in that they do not simply indicate lexical co-occurrences. Instead, there is a many-to-one mapping between possible surface forms and these more abstract features.

3.2.1 Word co-occurrences

The first class of features is simply occurrences of different words in the clauses related by the connective. Similar word co-occurrences have previously been shown to be useful for discourse level learning tasks such as inferring rhetorical relations (Marcu and Echihabi, 2002) and temporal relations (Lapata and Lascarides, 2004). In our case, we record which clause each word occurs in. This can be the clause immediately following the connective, i.e. the subordinate clause or the second of coordinated clauses. Alternatively it can be the clause that does not immediately follow the connective, i.e. the main clause in the case of a subordinating conjunction, or the first of coordinated clauses. We will use the subscripts $\perp$ and $\top$ to denote word occurrences in these two clauses, respectively, for example from (3.45) we obtain the set of co-occurrences for *whereas* shown in (3.46). (Technically we use multisets, i.e. repetitions of a word lead to multiple co-occurrences.)

(3.45) The cats ate meat, *whereas* the cows ate grass.

(3.46) $\{$ <whereas,the$_\top$>, <whereas,cats$_\top$>, <whereas,ate$_\top$>, <whereas,meat$_\top$>,
<whereas,the$_\perp$>, <whereas,cows$_\perp$>, <whereas,ate$_\perp$>, <whereas,grass$_\perp$> $\}$

Word co-occurrences were lemmatised using a stemmer based on Porter's (1980) suffix stripping algorithm. Lemmas with a frequency below 1000 per million in the BNC were then excluded. Finally, co-occurrences were indexed by their part of speech (POS) tags. These were readily available since we had already parsed each sentence in order to determine if it contained a discourse connective. The POS tags returned by the parser were those used in the Penn Treebank (Marcus et al., 1993). However we first clustered the part of speech tags, in part

Original Penn Treebank labels		New label	(description)
VB, VBD, VBG, VBN, VBP, VBZ	↦	VB	(main verbs)
NN, NNS, NNP	↦	NN	(nouns)
JJ, JJR, JJS	↦	JJ	(adjectives)
RB, RBR, RBS	↦	RB	(adverbs)
AUX, AUXG, MD	↦	AUX	(auxiliary verbs)
PRP, PRP$	↦	PRP	(pronouns)
IN	↦	IN	(prepositions)

Table 3.6: Clustering of POS labels

to remove information that was not specific to the lemma. The supercategories used are shown in Table 3.6. As a result, the final set of co-occurrences obtained from (3.45), repeated below, actually have the form shown in (3.48).

(3.47) The cats ate meat, *whereas* the cows ate grass.

(3.48) $\{ <\text{whereas,NN:cat}_\top>, <\text{whereas,VB:eat}_\top>, <\text{whereas,NN:meat}_\top>,$
$<\text{whereas,NN:cow}_\bot>, <\text{whereas,VB:eat}_\bot>, <\text{whereas,NN:grass}_\bot> \}$

The reason for recording part of speech information was so that the effects of different word classes could be studied. We now briefly discuss why various word classes might exhibit statistical co-occurrence relationships with discourse connectives.

Verbs Verbs often introduce the main predicate of a clause. As such they can play a crucial role in determining the appropriateness of different discourse connectives (Asher and Lascarides, 2003). For example, in (3.49) the contrast relation is set up by the conflicting verbs *love* and *hate*.

(3.49) Pat **loves** squash, *whereas* Sandy **hates** it.

Verbs are also important in descriptions of common sequences of events, also known as "scripts". For example *paying* often follows *eating*, as in (3.50).

(3.50) *After* John had **eaten** his meal, he **paid** for it.

This relationship between verbs and temporal connectives has been found to be strong enough to aid in the inferral of temporal connectives (Lapata and Lascarides, 2004). If we want to explore the use of co-occurrences with verbs for classifying discourse connectives, we just need to select the co-occurrences indexed by VB. For example, from (3.48) we extract the following co-occurrences:

(3.51) $\{ <\text{whereas,VB:eat}_{\top}>, <\text{whereas,VB:eat}_{\bot}> \}$

Nouns By indicating the participants of events, nouns also play an important role in the clause. They can thus also help set up contrast relations, as in (3.52).[5]

(3.52) Pat loves **squash**, *whereas* Sandy likes **tennis**.

In copular constructions, nouns can also introduce the main predicate of a clause, and the ability of some nouns to refer to events can also make them important to the discourse structure.

Adverbs Adverbs can introduce important modal information into a clause, e.g. *not, possibly* or *definitely*. Such modal adverbs can emphasise logical relations introduced by discourse connectives, as illustrated by (3.53).

(3.53) *If* I'm not on call, I'll **definitely** come to your party.

In addition, many adverbs also signal discourse relations, and so can be indicative of the discourse context in which a connective appears. This shall be discussed at more length in Section 3.2.2.

Auxiliary verbs Auxiliary verbs convey important information about tense and aspect, as well as voice. As such, they can be expected to show different co-occurrence patterns with different temporal connectives.

Adjectives, prepositions and pronouns There is probably less reason to expect adjectives, prepositions and pronouns to have informative co-occurrence patterns with discourse connectives. Nevertheless, adjectives can introduce predicates that set up contrast relations (e.g. *John is **tall**,* but *Jane is **short***), prepositions such as *before* and *after* can signal temporal relations,

[5]Note that *squash* is ambiguous, but that the contrast relation signalled by *whereas*, in combination with the noun *tennis*, helps to disambiguate it (Asher and Lascarides, 1995). However our model does not disambiguate word senses.

and pronouns can be indicative of the degree of subjectivity of connectives (Bestgen et al., 2003).

3.2.2 Co-occurrences with discourse markers

The second class of features indicates the occurrence of other discourse markers, including both structural connectives and adverbials, within the clauses related by a connective. Hirschberg and Litman (1993) have observed that discourse markers are likely to co-occur in the same sentence. Specifically, they find that if a phrase *X* that may or may not be a discourse marker is preceded by a discourse marker, then the likelihood of *X* being a discourse marker is increased. Discourse adverbials can occur in either of the clauses related by a connective, as illustrated by (3.54).

(3.54) **At first** they might be offended *but* **afterwards** they'd see I'd done them a service.

In addition, a pair of clauses related by a structural connective can also be introduced by another structural connective, as in (3.55) (adapted from example (36) of Webber et al. (2003)).

(3.55) John ordered three cases of the '97 Barolo, **but** he had to cancel the order *because* he then discovered he was broke.

As with co-occurrences with other classes of words, we will use the subscripts $\top$ and $\bot$ to indicate the clause in which a discourse marker co-occurs with a connective. For example, (3.54) and (3.55) produce the sets of discourse marker co-occurrences shown in (3.56) and (3.57), respectively.

(3.56) $\{ <\text{but,at first}_{\top}>, <\text{but,afterwards}_{\bot}> \}$

(3.57) $\{ <\text{because,but}_{\top}>, <\text{but,because}_{\bot}>\}$

There are several differences here with the features representing co-occurrences with words introduced above. Firstly, only a subset of words (or none at all) in each clause are discourse markers, so fewer discourse marker co-occurrences are stored than word co-occurrences. (As a consequence, co-occurrences of this sort will most of the time not be useful for applications involving disambiguating individual tokens, e.g. discourse parsing.) Secondly, discourse markers can be multiword expressions, as in *at first* in (3.56). Thirdly, no POS information is recorded with the discourse marker features. Instead, discourse markers are essentially treated as a lexical category in their own right.

Halliday and Hasan (1976, p. 237–238) were perhaps the first to note that when two discourse markers occur in the same sentence either they may be signalling different coherence relations, or they may both be signalling the same relation. In either case, the co-occurrence of a connective with another discourse marker provides information about the discourse context in which a connective appears. In the remainder of this section we aim to describe this relationship more fully. Webber et al. (2003) describe four different situations that can arise when a discourse adverbial appears in a clause that follows a discourse connective. The four cases are exemplified by (3.58)–(3.61).

In (3.58), both *because* and *then* relate the clause following *because*, i.e. the discovering, to other events. The discovering explains the cancelling event, and it also takes place after the ordering event.

(3.58) a. John loves Barolo.
b. So he ordered three cases of the '97.
c. But he had to cancel the order
d. *because* he *then* discovered he was broke.

In (3.59), the situation is slightly different, as the cause is not expressed by the clause following *because*. The reason why you should stop is instead the conditional structure introduced by *otherwise*, namely "if you don't stop you'll get a ticket".

(3.59) If the light is red, stop, *because otherwise* you'll get a ticket.

The third situation arises with adverbials signalling exemplification, such as *for instance* and *for example*. In (3.60), the not returning things is not why you shouldn't trust John, but just an example of why you shouldn't.

(3.60) You shouldn't trust John *because, for example,* he never returns what he borrows.

The final situation arises with adverbials such as *nevertheless* that signal that an underlying defeasible rule has been defeated. In (3.61), the defeasible rule incorporates the relation signalled by the temporal connective *while*. That is, it is presupposed that one does not normally think about fish while discussing politics.

(3.61) John is discussing politics *while* he is *nevertheless* thinking about fish.

In the first and second of these cases, it is also possible to recognise subcases on the basis of the location of the anaphoric argument of the discourse adverbial. In particular, the anaphoric argument may be the same as the (other) argument of the discourse connective, as in (3.62). Here, the adverbial *yet* indicates a contrast relation that is also signalled by the connective *but*.

(3.62) The pores in the skin are a classic example: they cannot become perceptible to us by themselves, *but yet* their presence in the skin can be deduced from sweat. (BNC)

In this case, the two discourse markers seem to be signalling the same discourse relation, however, this need not be the case. For example, in (3.63) the connective *but* seems to signal a Denial of Expectation, whereas *then* signals temporal succession between the same events.

(3.63) In one form of the myth he was killed and dismembered, *but then* his head floated. (BNC)

Although the connective and the adverbial need not signal the same discourse relation, it is necessary that they signal compatible relations, as otherwise incoherence results, as in (3.64) ("%" is used to indicate that a text is incoherent, whereas "*" is used to denote ungrammaticality).

(3.64) %Pat went to the delicatessen *before* he bought the paper *beforehand*.

In this case there is a clash between the temporal orderings being signalled. However the incompatibility can also arise through different orders of causation being signalled. Consider the following:

(3.65) The Greenhouse Effect accelerated *because* people used their air conditioners more.

(3.66) The Greenhouse Effect accelerated. *As a result* people used their air conditioners more.

(3.67) *The Greenhouse Effect accelerated *because as a result* people used their air conditioners more.

Even though both orders of causation are possible here (due to their being a positive feedback mechanism), it does not seem possible to use discourse markers to signal both directions of causation at once. Finally, although it is often possible to doubly signal a relation (e.g. the co-occurrence *but nevertheless* is quite common), there are constraints on this too. For example, consider that *suppose that* can be used to paraphrase *if*, but that both cannot be used simultaneously:

(3.68) *If* they are travelling at about sixty miles an hour, then they will arrive in twenty minutes.

(3.69) *Suppose that* they are travelling at about sixty miles an hour. Then they will arrive in twenty minutes.

(3.70) **If suppose that* they are travelling at about sixty miles an hour, then they will arrive in twenty minutes.

Co-occurrence constraints might even exist for the connective *and*, which can signal the widest variety of relations. Blakemore and Carston (2005) point out that the inferential (PRAGMATIC) interpretation of *so* seems unavailable when it co-occurs with *and*:

(3.71) These are his footprints; *so* he's been here recently.

(3.72) ?? These are his footprints; *and so* he's been here recently. ["??" is Blakemore and Carston's judgement]

In addition Blakemore and Carston argue that *after all* is unable to co-occur with *and* in cases where the former is used to signal evidence for an assumption for which no prior evidence has been given, as in:

(3.73) Let's start now; *after all*, we do want to finish before 6:00pm.

(3.74) ? Let's start now; *and after all*, we do want to finish before 6:00pm. ["?" is Blakemore and Carston's judgement]

Such preferences for or restrictions on how neighbouring discourse markers may be interpreted may result in statistical tendencies for certain co-occurrences to be more or less common.

A second case arises when the anaphoric argument to the adverbial is earlier in the discourse, as in (3.58), repeated in (3.75). Here the adverbial *then* indicates a sequence relation with the ordering event in (3.75b).

(3.75) a. John loves Barolo.
b. So he ordered three cases of the '97.
c. But he had to cancel the order
d. *because* he *then* discovered he was broke.

Sentence:	I like squash, but Bill doesn't.
Features:	POSITION=POST, EMBEDDING=1, NEG-VERB$_{\perp}$, VP-ELLIPSIS$_{\perp}$, PRONOUN$^{1}_{\top}$, STRUCTURAL-SKELETON$_{\top}$=NP-VB-NP, STRUCTURAL-SKELETON$_{\perp}$=NP-VB, ARGS$_{\top}$=SUBJ-OBJ, ARGS$_{\perp}$=SUBJ, WORDS$_{\top}$=3, WORDS$_{\perp}$=2, NPS$_{\top}$=2, NPS$_{\perp}$=1, PPS$_{\top}$=0, PPS$_{\perp}$=0, CLAUSES$_{\top}$=0, CLAUSES$_{\perp}$=0, MODALITY=<NULL,NULL>, MOOD=<DECL,DECL>, PERFECT=<NO,NO>, PROGRESSIVE=<NO,NO>, TENSE=<PRESENT,PRESENT>

Figure 3.5: Abstract features for an example sentence

When the discourse adverbial is incompatible with the structural connective, as, for example, with *and* and *beforehand*, then the adverbial's anaphoric argument must be previous to the discourse, as in (3.76), in which the feeling of confidence is before the surgery.

(3.76) I also had an upper Endoscopy... My consultant surgeon has an excellent record with this surgery *and beforehand* I did feel confident in putting my trust in him and God. (`http://www.geocities.com/lapro_fundo/andrew.html`, 27 March 2005.)

Regarding co-occurrences of discourse markers, it has also been suggested that markers that are less specific (i.e. can signal a wider variety of relations) are likely to precede markers that are more specific (i.e. can signal fewer relations) (Oates, 2000). However this may be simply because discourse connectives are, in general, less specific than discourse adverbials. For example, in Knott's (1996) taxonomy of about 150 discourse markers there are no connectives that are hyponyms of adverbials, although the converse relation often holds.

3.2.3 Abstract linguistic features

The third class of features we introduce include a range of features representing a variety of syntactic, semantic, and other information about the clauses related by a discourse connective. Figure 3.5 lists which of these abstract features (explained below) are present for a simple example. The features in this class are all extracted through automatic analysis of the parse trees. The complexity of the analysis required, for example analysing chains of auxiliary verbs for aspectual information, makes these features in this final class the most complex. There are

Polarity	Structural	Clause size	Specific lexical items
NEG-SUBJ	POSITION	WORDS	DO
NEG-VERB	EMBEDDING	NPS	BE
NPI-AND-NEG	STRUCTURAL SKELETON	PPS	TEMPEX
NPI-WO-NEG	VP-ELLIPSIS	CLAUSES	PRONOUNS

Figure 3.6: Summary of one dimensional abstract features

two main subtypes of these features, which have either one or two dimensional representations. We discuss each of these subtypes of turn.

One dimensional features Two of the one dimensional features recorded the location of the discourse connective within the clause. The POSITION feature took two values, indicating whether a discourse connective occurs between the clauses it relates, or prior to both. The distinction is illustrated by (3.77) and (3.78).

(3.77) I know the clerks by name; they answer me by mine. I say hello to a couple of them *before* I race to the opposite end of the building.

(3.78) "I'm taking my clothes," I say slowly, automatically, sadly, and with fear. *Before* my words are out, it's over.

Obviously this feature shows no variation for coordinating conjunctions. For subordinate conjunctions, however, the choice of whether to place the subordinate clause before or after the main clause often relates to the information structure of the sentence. That is, the clause containing old or given information is more likely to occur first (Heinamaki, 1972; Hamann, 1989; Schilder and Tenbrink, 2002), and reversing the order can cause problems for the reader, as illustrated by (3.79) and (3.80).

(3.79) I know the clerks by name; they answer me by mine. *Before* I race to the opposite end of the building, I say hello to a couple of them.

(3.80) "I'm taking my clothes," I say slowly, automatically, sadly, and with fear. It's over *before* my words are out,

Deep syntactic embedding of a discourse connective can indicate the subordination of multiple discourse relations and make processing more difficult, as in (3.81).

(3.81) If, *after* leaving the party, Jane drives home, she'll be lucky not to be done for drink driving.

The EMBEDDING features indicate the level of embedding, in number of clauses, of the discourse connective beneath the sentence's highest level clause. In doing this, we treat each connective as introducing a new level of embedding. For example, in (3.81) *if* is considered to be embedded one clause beneath the main clause, while *after* is embedded two clauses deep.

The remaining features recorded the presence of linguistic features that are localised to a particular clause. Like the lexical co-occurrence features, these were indexed by the clause they occurred in: either $\top$ or $\bot$.

Negation can play an important role in signalling contrast relations, as in (3.82).

(3.82) Pat likes squash, *whereas* Kim does**n't**.

We used two features to represent negation: NEG-SUBJ and NEG-VERB indicated the presence of subject negation (e.g. *nothing*) or verbal negation (e.g. *n't*), respectively. Position within the syntactic parse trees was taken into account when determining these features, so, for example, if the noun *nothing* was in object position then the feature NEG-SUBJ was not triggered.

Some discourse connectives are known to license Negative Polarity Items (NPIs) (Sánchez Valenzia et al., 1993). NPIs are a range of linguistic items such as *any* or *ever* which can occur in negated sentences, but cannot typically occur in non-negated sentences, as illustrated by (3.83) and (3.84). One discourse connective that can license NPIs in non-negated sentences is *before*, as illustrated by (3.85).

(3.83) *I ate any cake.

(3.84) I didn't eat any cake.

(3.85) The waiter cleared the table *before* I had eaten any cake.

The features NPI-AND-NEG and NPI-WO-NEG indicated whether an NPI occurred in a clause with or without verbal or subject negation.

Eventualities can be placed or ordered in time using not just discourse markers but also temporal expressions. The feature TEMPEX recorded the number of temporal expressions in each clause, as returned by a temporal expression tagger (Mani and Wilson, 2000).

If the main verb was an inflection of *to be* or *to do* we recorded this using the features BE and DO. Our motivation was to capture any correlation of these verbs with states and events respectively.

If the final verb of a clause is a modal auxiliary, this ellipsis of the main verb is evidence of strong cohesion in the text (Halliday and Hasan, 1976). It has also been argued that the resolution of such VP ellipsis is closely related to the coherence relation that relates the clause containing the ellipsis to the previous text (Kehler, 2002). We recorded this with the feature VP-ELLIPSIS.

Another indicator of textual cohesion is pronouns, and proportions of first and third person pronouns have been found to correlate with the degree of subjectivity of Dutch causal connectives (Bestgen et al., 2003). A class of features PRONOUNS^X represented pronouns, with X denoting either 1st person, 2nd person, or 3rd person animate, inanimate or plural.

The syntactic structure of each clause was captured using two features, one finer grained and one coarser grained. STRUCTURAL-SKELETON identified the major constituents under the S or VP nodes, e.g. a simple double object construction gives "NP VB NP NP". ARGS identified whether the clause contained an (overt) object, an (overt) subject, or both, or neither.

The length or size of a syntactic constituents can affect how they are arranged. For example, in English there is a preference for putting large constituents later in the sentence. It is conceivable that similar considerations work at the discourse level. The overall size of a clause was represented using four features. WORDS, NPS and PPS recorded the numbers of words, NPs and PPs in a clause (not counting embedded clauses). The feature CLAUSES counted the number of clauses embedded beneath a clause. This last feature can be indicative of their being discourse relations subordinate to the one signalled by the connective in question.

Two dimensional features This last class of features recorded combinations of linguistic attributes across the two clauses related by the discourse marker. In this case, attributes belonged to the clause as a whole, for which one of a fixed set of values exists for any (tensed) clause. For example, the MOOD attribute must take a value from the set {declarative, imperative, interrogative} for any clause. The choice of a two-dimensional representation of these features meant that combinations of attributes across the clauses could be represented. For example the MOOD feature would take the value $<\text{DECL}_\top,\text{IMP}_\bot>$ for the sentence shown in (3.86), and $<\text{INTERR}_\top,\text{DECL}_\bot>$ for (3.87).

(3.86) John is coming, but don't tell anyone!

Attribute	Possible values
MODALITY	FUTURE, ABILITY or NULL
MOOD	DECLARATIVE, IMPERATIVE or INTERROGATIVE
PERFECT	YES or NO
PROGRESSIVE	YES or NO
TENSE	PAST or PRESENT

Figure 3.7: Summary of two dimensional abstract features

(3.87) If John is coming, then how is he going to get here?

Blakemore and Carston (2005) observe that even the most general discourse connective *and* shows curious restrictions in its ability to conjoin declarative clauses with ones of other types:

(3.88) I went to the lecture and who do you think I saw?

(3.89) %I went to the lecture and who was there?

(3.90) Your mother has already left. Go home!

(3.91) %Your mother has already left and go home!

The attributes used to construct these two dimensional features were MOOD, MODALITY, PERFECT, PROGRESSIVE and TENSE, and the possible values that each attribute could take are shown in Figure 3.7. A distinction is often drawn between discourse relations which hold between the semantic contents of clauses, and those which hold at a level of pragmatics or speech acts (Halliday and Hasan, 1976; Sweetser, 1990; Martin, 1992; Sanders et al., 1992; Knott, 2001). If a clause is interrogative or imperative, then the relation can only hold at the "pragmatic" level. The MOOD feature we use therefore relates to this distinction.

Tense and aspect can have quite subtle effects on the determination of the nature of the temporal relation that holds between events (Moens and Steedman, 1988; Glasbey, 1995). Our TENSE, PERFECT and PROGRESSIVE features aim to capture any trends that may hold between tense and aspect and the various discourse connectives.

Finally, certain connectives, such as conditionals, can alter the modality of the clauses they relate. We therefore use the MODALITY feature in order to represent explicit signalling by the auxiliary verbs of different modalities.

after	although	and	as [sic]
as long as	as soon as	as(1)	as(2)
as(3)	assuming that	because	but
considering that	even if	even though	ever since
for	given that	however	if
if ever	if only	in case	in order that
in that	insofar as	just as	now
now that	on condition that	on the assumption that	on the grounds that
once	or	or else	or rather
provided that	seeing as	since	so
so that	supposing that	the instant	the moment
the way	then(1)	though	to the extent that
unless	until	when	whereas
while(1)	while(2)	yet	

Table 3.7: Structural connectives in Knott's (1996) taxonomy. Integers represent sense numbers.

Many of the abstract features discussed above cannot always be identified automatically. Because of this, and also because of the inevitability of parsing errors, there is always a degree of noise in the feature counts. When implementing the heuristics for identifying each of the abstract features, a conservative approach was taken. For example, a marked feature such as the imperative mood was only signalled as such if the syntax strongly suggested this was case, and short lists of unambiguous lexical items were used to identify negation, NPIs and pronouns.

3.3 A taxonomy of discourse connectives

We now discuss the creation of a gold standard taxonomy for evaluating the experiments in Chapters 5 and 6. Knott (1996) presents a taxonomy which represents substitutability relationships between 152 discourse markers. Of these, 97 are discourse adverbials and 55 are structural connectives; the latter are listed in Table 3.7. By extracting just these structural dis-

before	but not after	but not when	but then
by the time	despite the fact that	else	even after
even before	even though	even when	except
except after	except before	except since	except when
for fear that	for the reason that	in the hope that	lest
much as	notwithstanding that	only after	only before
only if	only until	only when	presumably because
until after	whether or not	which is why	which was why

Table 3.8: The connectives added manually to Knott's taxonomy

course connectives from the taxonomy, along with the relationships that hold between them, we obtained a taxonomy consisting solely of structural connectives. Then, using Knott's Test for Substitutability (Figure 2.13), this taxonomy was extended manually by adding the 32 connectives shown in Table 3.8. Instead of presenting in detail the tests for substitution that we carried out, we instead discuss here a number of practical issues which arose in applying Knott's substitution methodology. (Details concerning where the new connectives were inserted into Knott's taxonomy can be found Appendix A.)

The first practical issue arose directly from the definitions of the substitutability relationships. Observe that the definitions of SYNONYMY, HYPONYMY, and EXCLUSIVE are framed in terms of generalisations over all possible contexts in which a discourse marker appears. As Knott points out (in Appendix B, page 172), such claims cannot be verified, only falsified. The problem is precisely what Popper (1959) calls "The Problem of Induction" (typically illustrated via reference to white swans). It is therefore necessary for the practical analyst to at some point place faith in the belief that he or she has considered all relevant types of contexts that a connective appears in, so that he or she can then decide which substitutability relationship holds. This raises the question of how much empirical evidence should be required to support such a belief. In an ideal world, the analyst would be able to enumerate all possible types of context, abstracting over differences between individual contexts which cannot affect the appropriateness of discourse connectives. For example, (3.92) and (3.93) have very similar semantic structures, and there is no connective that is suitable in one of them but not the other.

(3.92) Pat likes skiing. Chris prefers snowboarding.

(3.93) Sandy likes Thai food. Lee prefers Japanese.

If such an enumeration were possible, then only a single example of each type of context would need to be tested. However in practice it is not possible to partition contexts into discrete classes such that all and only aspects relevant to the suitability of discourse connectives are captured. (Indeed, such a classification would be dependent on a theory of what discourse connectives mean, leading to a circularity in the research methodology.) All the relevant aspects of discourse are simply not known. Another option would be to take some number N of substitutability tests to be sufficient evidence for deciding on a substitutability relationship between connectives. But this raises the problematic question of how many tests should be required. For example, if carrying out three substitution tests suggests SYNONYMY, should that be taken as good enough proof? How about if ten tests are carried out? Or fifty?

Popper argues that hypotheses cannot be verified, but they can be "corroborated" to different degrees. Furthermore, the degree of corroboration of a hypothesis increases along with the severity of the tests to which the hypothesis is subjected. One possible avenue for sharpening Knott's methodology is the incorporation of inferential statistics such as are commonly used in social sciences like psychology. In particular, it would be useful to be able to estimate the probability of making a Type I Error, i.e. the probability of accepting that a given generalisation over all contexts holds, when in fact it does not. We now briefly discuss what would be required for the use of such inferential statistics.

To begin with, note that there is presumably only a finite set of aspects of the linguistic context that are relevant to the appropriateness of discourse markers. These are known to include temporal structure, event structure, and negation, and would presumably preclude the choice of particular referring expressions, as well as various aspects of global discourse structure that do not affect local decisions about the signalling of relations. The finiteness of this set follows from the fact that the number of discourse markers is itself finite, along with an assumption that linguistic features are discrete, as opposed to forming a continuum. However, as mentioned above, the precise content of this set of relevant aspects of context is not known. We therefore proceed by treating these as hidden variables.

Suppose an analyst is asked whether one discourse marker *after* can always be substituted for another one *when*. Further, suppose the analyst is linguistically competent, but extremely naive in that they have no prior intuitions about the range of contexts in which *when* can be used. They might therefore randomly select a collection of N texts containing *when* and proceed to judge whether *after* can be substituted for *when* in each case. Suppose that for each of the N

texts it can indeed be substituted. The null hypothesis in this case is that *after* cannot always be substituted for *when*. We can exclude the possibility that *after* can never be substituted for *when*, so the null hypothesis is effectively that *after* can only sometimes be substituted for *when*. In order to apply an inferential statistic, we must now assume that we can estimate the following conditional probability:

$$P(X \text{ is not substitutable for } Y \text{ in a particular context,}$$
$$\text{given that } X \text{ is CONTINGENTLY SUBSTITUTABLE for } Y)$$

The precise value of this probability is difficult to estimate, however in practice a lower bound on this value can be used to ensure conservatism in rejecting the null hypothesis. Such a lower bound might be obtained empirically, via repeatedly carrying out Knott's Test for Substitutability with pairs of connectives known to be CONTINGENTLY SUBSTITUTABLE. We could then use the sign test (based on the binomial distribution) to calculate the probability of making a Type I error.

Now suppose that we present the same task to another analyst, equally competent, but less naive than the first. Suppose that this analyst recognises that *when* sometimes signals temporal overlap, and sometimes temporal succession, but has no prior intuitions regarding the contexts in which it signals one relation rather than the other. In this case a more sophisticated methodology is appropriate. The analyst might take N examples where temporal succession is signalled, and another N cases involving temporal overlap. Imagine that *after* can be substituted in each of these $2N$ examples (although it obviously cannot!), then an inferential statistic with more sophistication than the simple sign test would be required to estimate the probability of the null hypothesis.

To sum up, we have suggested how inferential statistics might be incorporated to make the substitution methodology more rigorous. However pursuing this matter further is beyond the scope of this thesis.

The second practical issue that arose in applying the substitution methodology is the sampling from corpora of example sentences containing the connectives. This has already been alluded to above, where we suggested that if an analyst knows that a discourse marker has different usages, then each of these cases should be considered. The decision of which corpus to use also affects the sampling; we obviously want a corpus which contains all the relevant different usages of a discourse marker.

In our case, we obtained example sentences by searching the web for the connective, and then examining the top twenty or so hits for a range of different usages of the connective.

Usages that were felt intuitively to be different in a relevant fashion were extracted and later used for making substitutability judgements. In using the web as source of linguistic data, we could potentially access a large range of genres and styles. However the web has no quality control mechanism, and the identity of the author of a text, and whether they are a native speaker of English, is often unknown. This can be problematic when, as happened in a few cases, we personally found the use of a connective in a text to be unacceptable. This raises the issue of speaker variability, as well as the question of whose judgements the taxonomy we were constructing was meant to represent.

We decided that the taxonomy was not meant to represent an individual's perspective, but something more general. Although we usually only have access to our own judgements on acceptability, we treated cases which we personally found unacceptable as rare insights into how we might differ from the general population. That is, we gave preference to the empirical fact that these other usages do get produced, over our introspective judgements on acceptability. Of course, this cuts right to the heart of issues involving the use of competence data versus performance data. Unfortunately, little work has been done on measuring inter-speaker agreement on the acceptability of discourse markers.

Related to this, another practical consideration was what to do if we disagreed with any of Knott's judgements on substitutability. We decided to treat Knott's taxonomy as infallible (some minor inconsistencies in Knott's taxonomy were resolved after personal communication with Knott; these are detailed in Appendix A). The reason for this was that it is a useful experimental procedure to use a resource that is freely available to all. If we had started modifying Knott's taxonomy as we saw fit, the usefulness of having a standard, published, and widely available resource would have been eroded.

The fifth practical consideration concerned the question of which connectives to use as candidate substitutions. Given the number of connectives, and the number of example texts, making all pairwise comparisons for each connective and each text was impossible. Instead, in each case we chose a subset of connectives which were intuitively felt to be related to the original connective in some way.

Our decision to compare pairs of intuitively related connectives was closely related to Knott's presentation of his taxonomy. The taxonomy is organised into ten "categories", such as "temporal phrases", "result phrases", and so on. One useful aspect of using these categories is that many EXCLUSIVE relationships can be succinctly expressed. For example, *because* is listed as an "exclusive cause phrase" (i.e. it always signals a cause), whereas *meanwhile* is an "exclusive temporal phrase". It follows directly that *because* and *meanwhile* are EXCLUSIVE.

Another practical decision to be made concerned which version of the Test for Substitutability to use. The test is originally introduced in a succinct form by Knott and Dale (1994), however this original formulation is elaborated by Knott (1996) with the following modifications:

- The Test can succeed if the punctuation needs to be altered to make the new discourse marker acceptable.
- The Test can succeed if adding "additional or alternative" discourse markers to other clauses in the text makes the new discourse marker acceptable.
- The Test should overlook stylistic differences.
- The Test should overlook differences resulting from different sizes of text spans.
- The Test should disregard the amount of background knowledge the reader is assumed to possess.

It is this 1996 version of the Test that we presented in Figure 2.13. Knott and Sanders (1998) then modify the Test further by allowing discourse markers to be considered substitutable even if they occur in different clauses. For example, *so* is considered substitutable for *because* in the context (3.94) because of the acceptability of (3.95).

(3.94) *Because* Jane liked sailing boats, she took a job with a charter company.

(3.95) Jane liked sailing boats, *so* she took a job with a charter company.

Knott and Sanders use the term "swap-substitutable" to describe such cases.

In order to maintain compatibility with Knott's (1996) taxonomy, we used the 1996 version of his Test for Substitutability. However, if one was starting from scratch then the decision of which version to use might not be so clearcut.

The last practical issue concerned the categorical nature of the acceptability judgements. A particular substitution might be felt to be acceptable, but only barely, while another might feel unacceptable, but only just. This is directly comparable to the problem of making categorical judgements as to the grammaticality of sentences (Sorace and Keller, 2005). In general, a theory of language which can account for quantitative differences in acceptability ratings, rather than just binary judgements, might be considered superior to one that cannot. However the adopted methodology did not account for such differences. One could imagine taking a poll in

such borderline cases, to determine whether the majority of speakers felt the substitution to be acceptable or not. However it was felt that relying on the judgements of an individual analyst was more in keeping with the spirit of Knott's approach, and so this was the approach that was adapted.

To summarise, many of the practical issues that arose do not pertain solely to the particular task at hand. Instead, they relate to many of the major meta-theoretical issues of linguistics, such as what type of data should be used, how that data should be judged, how differences in judgements should be resolved, and the validity of making linguistic generalisations. The latter has particular relevance to the substitution task, and we have outlined what would be required to develop a more rigorous approach using inferential statistics.

3.4 Summary

In this chapter we have outlined the three main data requirements for the experiments in the coming chapters. These are: 1) a collection of texts containing discourse connectives, and 2) a set of features for representing the contexts in which discourse connectives appear, and 3) a taxonomy representing substitutability relationships between discourse connective that will be provide a gold standard for evaluation. We then met these requirements by describing a new methodology for mining example sentences from the web, presenting features for representing context, and discussing the manual extension of Knott's taxonomy. In combination, these allow us to produce representations of the distributions of discourse connectives. In the following chapter, we use these distributions to acquire attributes of discourse connectives, and in Chapter 5 we use these distributions to learn substitutability relationships between discourse connectives.

Chapter 4

Acquiring knowledge about individual connectives

This chapter presents a series of experiments into automatically learning attributes of discourse connectives. The automatic acquisition of these attributes is important for a number of reasons. Firstly, it enables the rapid classification of a large number of connectives with less human effort, making use of subtle distributional properties that might not be obvious to a human expert. Secondly, although here we classify connective types, this constitutes a first step towards the automatic classification of connective tokens, e.g. disambiguating polysemous discourse connectives, which is an important step in discourse parsing. Thirdly, this chapter contributes directly to one of the major aims of this thesis, which is to demonstrate that automatic techniques can be applied to the task of acquiring knowledge about discourse connectives. Somewhat more indirectly, it also provides some support for the hypothesis that connectives with similar meanings have similar distributions. This hypothesis is addressed in a direct fashion in the following chapter.

The experiments in this chapter concern learning attributes which specify different semantic aspects of discourse connectives. These attributes are grouped along four independent dimensions, and in combination they specify factors which are important for interpreting and reasoning about texts, such as the temporal ordering of events, the modality of propositions, and the presupposing of causal rules. The four dimensions of attributes will be discussed in detail later in the chapter, and for the time being just concise introductions with a few illustrative examples will be given.

The **polarity** dimension has the effect of distinguishing pairs of discourse connectives such

as *so* or *but*, illustrated in (4.1) and (4.2). While *so* introduces some kind of implication or cause, *but* signals a violation of a causal rule (at least in this example). We will describe *so* as having POSITIVE POLARITY, and *but* as having NEGATIVE POLARITY. (Note that, in line with our use of fonts in Chapter2, we will use SMALL CAPITALS for classes of relations, and **bold** for the dimensions on which we classify connectives.) The latter is related to various categories of relations discussed in Chapter 2, including Martin's (1992) CONTRAST and CONCESSION relations, Kehler's (2002) CONTRAST, VIOLATED EXPECTATION and DENIAL OF PREVENTER relations, Sanders et al.'s (1992) and Knott's (1996) NEGATIVE POLARITY primitives.

(4.1) Jim had just washed his car, *so* he wasn't keen on lending it to us. (Knott, 1996, p. 100)

(4.2) It was odd. Bob shouted very loudly, *but* nobody heard him. (Knott, 1996, p. 100)

The **veridicality** dimension indicates whether discourse connectives imply the truth of the clauses they connect. This is the case with *and* in (4.3), but not with *if* in (4.4).

(4.3) John is always gloomy *and* he never has anything interesting to say.

(4.4) You can stay up with us *if* you promise to be quiet. (Knott, 1996, p. 189)

The **type** of a connective indicates whether it explicitly signals an additive, temporal or causal relation. The distinction is illustrated by examples (4.5), (4.6) and (4.7), respectively.

(4.5) The Normans invaded Britain, *and* the Vikings did (too).

(4.6) The Normans invaded Britain *after* the Vikings did.

(4.7) The Normans invaded Britain *because* the Vikings did.

Finally, the dimension of **direction** distinguishes between different temporal orderings, as well as the different arguments to causal relations. For example, in (4.8) and (4.9) the eventuality introduced by the connective follows, temporally and causally respectively, the eventuality of the previous clause; the converse is true in (4.10) and (4.11).

(4.8) Jane left *before* it got dark.

(4.9) Sue was sick, *so* she stayed in bed all day.

(4.10) It got dark *after* Jane left.

(4.11) Sue stayed in bed all day *because* she was sick.

A discourse connective can be classified along more than one dimension, for example *although* is both CAUSAL and has NEGATIVE POLARITY. On the other hand, some dimensions are not appropriate for some connectives, e.g. **direction** is irrelevant for the connective *but*, but is relevant for the connective *after*.

The acquisition of semantic attributes will be viewed as a classification task, where the aim is to classify lexical items according to the presence or absence of the properties in question. This experimental paradigm has previously been applied to a wide range of lexical phenomena, including verb classes (Merlo and Stevenson, 2001), verb aspect (Siegel and McKeown, 2000) and noun countability (Baldwin and Bond, 2003). We proceed by discussing the attributes we will use in more detail. We then discuss the experimental setup, before presenting the experiments into learning attributes on each of the four dimensions.

4.1 Attributes to be learnt

In this section we relate the attributes to be acquired to the previous literature presented in Chapter 2. We also discuss the creation of the gold standard classifications used in the experiments. It is important to stress that in all these cases it is lexical *types*, not tokens, that are classified, and the same holds for our experiments. Naturally, the meaning conveyed by some discourse connectives can be augmented by their context, and for applied tasks such as discourse parsing disambiguating at the token level is important. Nevertheless, classifying connective types as to the invariant information they convey can be considered a prerequisite for classifying the meaning conveyed by connective tokens in context. As a result of these considerations, we omit ambiguous connectives from the experiments when the various senses differ along the dimension we wish to classify. For example, the two senses of *while* differ on the **polarity** dimension (the temporal sense is POSITIVE; the contrastive sense NEGATIVE), so *while* is omitted from the **polarity** experiment. Whether the meaning of a discourse connective is ambiguous or merely underspecified is not always obvious, however a few heuristics were used when creating the gold standards. If a connective has two meanings which are radically different, then it is treated as ambiguous. The literature was also taken into account. If it has been proposed that a connective belongs to adjacent (or otherwise closely related) subcategories, this suggests underspecification may be involved. As an example of this, consider Martin's (1992) network of simultaneous temporal relations shown in Figure 4.1. Here *when*

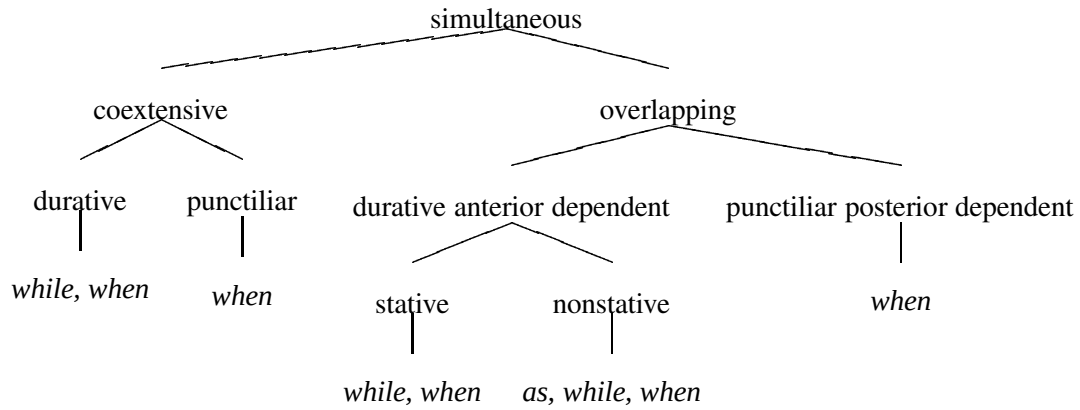


Figure 4.1: Martin's (1992) network of simultaneous temporal relations

occurs at every leaf node, which we take as evidence for *when*'s being underspecified with respect to the distinctions that Martin draws (the meaning of these distinctions need not concern us here).

4.1.1 The polarity dimension

The concept of discourse relations (and by extension discourse markers) having polarity is a recurrent one in the literature, although it is often conceived of in slightly different ways, as we will discuss below. These differences in attempts to provide a precise definition suggest that the concept is not a simple one; however it is widely agreed that prototypical discourse relations exhibiting NEGATIVE POLARITY include contrast, concession and denial of expectation. Discourse connectives commonly used to signal these types of relations include *but, even though* and *although*. We first examine the role of polarity in the literature before presenting the gold standard classification used in our experiment.

Halliday and Hasan (1976) use the terms "negation" and "polarity" to describe the relation between sets of discourse markers. For example, they describe *otherwise* as variously a "reversed polarity" conditional, and "the negative form of the conditional" (although they do concede that the latter description is misleading). In support of this analysis, they mention that *otherwise* can be paraphrased by *if not*, at least in certain contexts, as in (4.12). In other cases, where there is negation in the preceding context, *if so* is the correct paraphrase, as in (4.13).

(4.12) It's the way I like to work. One person and one line of enquiry at a time.

Otherwise/If not there's a muddle.
(Adapted from [5.50]a., of (Halliday and Hasan, 1976).)

(4.13) I was not informed. *Otherwise/If so/#If not* I should have taken some action.
(Adapted from [5.51], of (Halliday and Hasan, 1976).)

Similarly, *in this respect* is described as differing in polarity from *in other respects*, presumably because of the similarity between *other* and *not this*. Negation is mentioned explicitly in the description of *nor* as indicating a "negative additive" relation, and in the description of *by contrast* as indicating a "negative comparison". However the discourse connectives *but* and *though* are not described by Halliday and Hasan in terms of either negation or polarity, but are instead members of a class of "Adversative" discourse markers.

Martin (1992) uses the terms "negative" and "polarity", although he concedes that he does not use them with a constant meaning (p. 195). *Unless* is classified as a negative conditional on the grounds that *A unless B* is logically equivalent to *not A if and only if B*. On the other hand, the classification of *lest* as negative rests on the undesirability of the clause it introduces. Martin does not describe *but* and *although* in terms of negative polarity either, instead describing them as "contrastive" and "concessive".

In contrast to Martin, Sanders et al.'s (1992) classification of discourse relations places great importance on the concept of the negative polarity of discourse relations being a single phenomenon. For the first time, a formal definition of polarity is given. If there is a relation between the propositions expressed directly by two text segments, the relation is "positive". If instead the relation holds between one of the text segments and the negation of the other, the relation is "negative". For example, (4.14) is analysed as involving a causal relation with antecedent *not having any political experience* and consequent *not being elected president*, as shown in (4.15).

(4.14) *Although* he didn't have any political experience, he was elected president. (Sanders et al., 1992, (17)E, p. 10)

$$\neg(\text{have experience}) \Rightarrow \neg(\text{be elected}) \tag{4.15}$$

A similar effect can be obtained using the connective *but*, as in (4.16).

(4.16) He didn't have any political experience, *but* he was elected president.

These examples also illustrate that for Sanders et al. *but* and *although* are prototypical markers of negative discourse relations. In fact, this conception of negative polarity is closely related to Halliday and Hasan's paraphrasing of *otherwise* by *if not* and *if so* in (4.12) and (4.13). In (4.12), there is a conditional relation between the negation of working one at at time and there being a muddle:

$$\neg(\text{one at a time}) \Rightarrow \text{muddle} \tag{4.17}$$

In contrast, in (4.13), there is a conditional relation between the negation of not being informed and taking action. In this case, the two negations cancel, explaining why *if so* is the correct paraphrase in (4.13):

$$\neg(\neg\ \text{informed}) \Rightarrow \text{taken action} \quad \equiv \quad \text{informed} \Rightarrow \text{taken action} \tag{4.18}$$

Like Sanders et al., Knott (1996) is concerned with productivity and cognitive plausibility. Knott explicitly employs the notion of a defeasible rule (Hobbs et al., 1993). A relation is Positive if a) the defeasible rule relates the propositions conveyed by the text and b) this rule is upheld. Whereas it is Negative if both a) the consequent of the rule is inconsistent with the text, and b) the rule is defeated. Knott would therefore analyse (4.14) slightly differently to Sanders et al. For Knott, (4.14) involves an underlying rule that if *someone doesn't have political experience* then *they don't get elected*. However the rule is defeasible, and its defeat in this case is critical to the conclusion that the relation is NEGATIVE. Similarly, in (4.19) there is an underlying rule that if *you gave me a thousand pounds* then *I would vote for Major*, but the rule is defeated.

(4.19) I wouldn't vote for Major *even if* you gave me a thousand pounds. (Knott, 1996, p. 102)

Observe that if the reader does not previously know the political beliefs of the writer, then he or she must accommodate the defeasible rule.

As one final example, consider the use of *whereas* in (4.20).

(4.20) Pat plays tennis, *whereas* Chris plays squash.

For this to be appropriate, Chris mustn't play tennis. We therefore have play(pat,tennis) and ¬play(chris,tennis), and it is this usually taken as evidence for the relation being negative (Spooren, 1989).

Knott (1996, p. 106) has proposed an alternative analysis of cases like (4.20), however, whereby the relation is negative because of an inability to generalise from Pat's playing tennis to Chris' doing so. Knott's notion of polarity bears most similarity to that of Sanders et al., and he goes one step further in explicitly classifying the connectives *but* and *although* (rather than just the relations they can signal) as having negative polarity. Knott also introduces explicitly the notion of underspecification of polarity, for example *and* is described as being underspecified in this respect. This is because *and* is sometimes used when the defeasible rule is upheld, as in (4.21), and sometimes when it is defeated, as in (4.22).

(4.21) Jim has just washed his car, *and/%but* he wasn't keen on lending it to us. (Knott, 1996, p. 100)

(4.22) It was odd. Bob shouted very loudly, *and/but* nobody heard him. (Knott, 1996, p. 100)

Louwerse (2001) accepts Sanders et al.'s definition of polarity, adding that polarity can also be defined informally using the notion of "opposition". Louwerse presents eye-tracking data which shows negative discourse connectives receive more regressions than positive ones, suggesting this may be a psychologically real parameter. For Louwerse, *and* has positive polarity, in contrast to Knott, who considers *and*'s polarity to be underspecified. This is because for Louwerse *and* is associated with logical conjunction, which he associates with positive polarity. However we do not make this association, instead the issue of logical conjunction is captured by our dimension of **veridicality**, which is the subject of a separate experiment.

The notion of **polarity** we adopt is basically that of the closely related characterisations of Knott (1996) and Sanders et al. (1992): a discourse connective has NEGATIVE POLARITY if it signals an underlying rule relating the proposition expressed by one clause and the negation of the proposition expressed by the other. We simplify Knott's analysis slightly, in that we do not consider a separate class of markers with underspecified polarity. Instead, our gold classes, introduced below, indicate whether or not a discourse connective *always* signals a NEGATIVE POLARITY discourse relation. For example, every token of *but* involves NEGATIVE POLARITY (although we can remain agnostic as to which of Spooren's or Knott's analysis of (4.20) is more desirable), whereas the same is not true of *and*, hence for the purposes of our experiment *and* is classified as POSITIVE POLARITY.

POS-POL	NEG-POL
after, and, as, as soon as, because, before, considering that, ever since, for, given that, if, in case, in order that, in that, insofar as, now, now that, on the grounds that, once, seeing as, since, so, so that, the instant, the moment, then, to the extent that, when, whenever	although, but, even if, even though, even when, only if, only when, or, or else, though, unless, until, whereas, yet

Table 4.1: Discourse markers in the **polarity** experiment

4.1.2 The polarity gold standard

Experiments involving the classification of lexical types require a gold standard classification to evaluate against. The gold standard classes of POSITIVE and NEGATIVE discourse connectives used in this experiment are shown in Table 4.1. We noted above that *and* is classified as POSITIVE POLARITY because it does not always involve a NEGATIVE POLARITY relation. So more precise terms for our classes might be ALWAYS-NEGATIVE and NOT-ALWAYS-NEGATIVE, however for the sake of simplicity we adopt the conventional nomenclature. As a further example, consider that *while* does sometimes, but not always, signal a negative polarity relation. However in this case it is clear that there are two distinct senses: a temporal one, and a contrastive one. Because of this ambiguity, *while* is omitted from the experiment altogether.

The reasons for *until* being NEGATIVE POLARITY are not obvious, despite both Knott and Louwerse classifying it as such. Knott does not give explicit reasons for classifying *until* as NEGATIVE POLARITY, but a trace of his reasoning can perhaps be found in one of his substitutability tests, shown in (4.23) and (4.24). In (4.23), *Either... or* indicates NEGATIVE POLARITY, and the fact that *Until* can be substituted for *Either... or*, as in (4.24), suggests that *until* is NEGATIVE POLARITY too. However, this example on its own is not enough to force this conclusion, since it is also logically possible that *until* could be underspecified with respect to polarity.

(4.23) *Either* you settle the matter amicably, *or* you will never be friends again.

(4.24) *Until* you settle the matter amicably, you will never be friends again.

Louwerse explicitly addresses the classification of *until*. For him, *until* is NEGATIVE POLARITY because in *x until y* the situation *y* is related to the termination of situation *x*. He admits though

that this notion of negation is "a little different from that in [other] cases". For Louwerse, *until* is an exception to the rule that most TEMPORAL relations have POSITIVE POLARITY. It is presumably this temporal nature of *until* that leads to its omission from Sanders et al.'s (1992) discussion, as they purposefully avoid classifying temporal relations. We follow the literature by including *until* in the NEGATIVE POLARITY class, although we note that this classification does seem to be disputable.

4.1.3 The veridicality dimension

The second experiment concerns the acquisition of the property of veridicality. A discourse connective is veridical if it implies the truth of its two clauses. For example, *but* is veridical, so (4.25) implies that both John wasn't happy, and that Sue was happy.

(4.25) John wasn't happy, *but* Sue was.

In contrast, a connective like *if* is not veridical, so that in (4.26) neither John's nor Sue's happiness is implied.

(4.26) *If* Sue was happy, John wasn't.

Similarly, no inferences about the truth of the individual clauses can be made in (4.27) (although there is of course a relation *between* their truth values).

(4.27) *Either* John was happy, *or* Sue was.

The actual term "veridical" is not used in the major classifications of discourse markers. Various subclasses of non-veridical discourse markers are often distinguished though. For example, categories of ALTERNATIVE (e.g. *or else*) and CONDITIONAL (e.g. *if*) are used by Halliday and Hasan (1976), while Martin (1992) has closely related categories of ALTERNATION and CONDITION. In contrast, none of the parameters of Sanders et al.'s (1992) classification refer to the truth of the related propositions. Conditionals are not mentioned at all in their study. They do briefly consider disjunction, however, suggesting it can be analysed along the same lines as contrastive relations, following Longacre's (1983) claim that "While contrast turns on two points of difference, alternation turns on one point of difference."

Knott (1996) makes use of a MODAL STATUS parameter, taking values of either ACTUAL or HYPOTHETICAL. These are defined not in terms of truth values, but in terms of how much of the "preceding context" is known by the speaker (or writer). For example, in (4.28) the speaker knows that Mary will arrive home at some point, and so the modal status is ACTUAL.

(4.28) *When* Mary gets home, ask her to call me. (Knott, 1996, p. 121)

By contrast, in (4.29) the speaker does not know whether Mary will get home before eleven o'clock, and so the modal status is HYPOTHETICAL.

(4.29) *If* Mary gets home before eleven, ask her to call me.

For Knott, the disjunction *or* is also HYPOTHETICAL, whereas *before* is ACTUAL, although no reasons are given. However *before* also has a counterfactual use, as in:

(4.30) The children left *before* anyone arrived.

Despite veridicality not being used explicitly in classifications of discourse markers, it has been used in the more semantically-oriented literature on discourse connectives. Sánchez Valenzia et al. (1993) examine the relation between veridicality and the licensing of polarity-sensitive items such as *anything* or *need*. They show that veridicality relates to semantic monotonicity, and propose that this can explain the licensing behaviour of polarity sensitive items by certain connectives, as can be seen by contrasting (4.30) with (4.31).

(4.31) *Anyone arrived *after* the children left.

Tenbrink and Schilder (2001) analyse *before* as being non-veridical, on the basis of sentences like (4.32), where the clause following *before* is not implied to be true.

(4.32) Mary left the party *before* she punched anyone.

Indeed, the opposite is the case: it is implied (at least under one reading) that Mary did not in fact punch anyone. In other contexts, however, it can be implied that she did:

(4.33) A: Was it at the party that Mary began punching people?
B: No, Mary left the party *before* she punched anyone.

The concept of veridicality plays an important role in Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), which attempts to model the formal truth conditional aspects of discourse. The description of each discourse relation within SDRT indicates clearly whether the relation is veridical or not. A formal definition of veridicality in discourse is given within dynamic semantics.

In summary, although "veridicality" is not mentioned explicitly in much work on discourse connectives, it is a relatively intuitive and uncontroversial concept, with non-veridical discourse

VERIDICAL	NON-VERIDICAL
after, although, and, as, as soon as, because, but, considering that, even though, even when, ever since, for, given that, in order that, in that, insofar as, now, now that, on the grounds that, once, only when, seeing as, since, so, so that, the instant, the moment, then, though, to the extent that, until, when, whenever, whereas, while, yet	assuming that, even if, if, if ever, if only, in case, on condition that, on the assumption that, only if, or, or else, supposing that, unless

Table 4.2: Discourse markers in the **veridicality** experiment

relations having two main subtypes: alternation/disjunction and conditionals/hypotheticals. It is reasonable to propose that experiments should make a distinction between these subclasses, however in the experiment that follows we instead have just two classes: VERIDICAL and NON-VERIDICAL. There are three reasons for doing this. Firstly, the concept of veridicality is fairly straightforward, and from a methodological perspective it is sensible to try and model this basic distinction before attempting to learn finer-grained distinctions (for which data sparseness may be more of an issue). For experimental purposes, there is also the practical problem that only a few connectives signal disjunction, which would make experiments involving this class somewhat random and uninformative. Secondly, at a semantic level, disjunction does seem to be closely related to conditionals, since *A or B* implies *if not A, then B*. Thirdly, we believe that for many practical NLP applications, just knowing whether particular clauses in a text are implied to be true is of crucial importance. While some tasks such as Text Classification focus not on truth values but on "about-ness", for other tasks such as Summarisation or Question Answering, where clauses may be retrieved from a text and presented out of context, the truth values of propositions out of context is of central importance.

4.1.4 The veridicality gold standard

The gold standard classes of VERIDICAL and NON-VERIDICAL discourse connectives used in the experiment are shown in Table 4.1. The construction of the gold standard classes was more straightforward than it was for the **polarity** experiment, since the definition of veridicality is more straightforward, and less disputed. As a result, there are slightly more discourse connectives in this experiment; we evaluate on 49 connectives, compared to 43 in the **polarity**

experiment. We do however deliberately exclude *before* from the experiment, as its behaviour with regard to truth conditions is quite complex, as we saw above, in (4.32) and (4.33).

In our discussion of **polarity** we noted that *and* can often be used when a NEGATIVE POLARITY discourse relation is intended, but that for the purposes of our experiment we classified it as POSITIVE POLARITY because it did not always signal a negative relation. Taking a similar approach, one could argue that *before* should be classified as NON-VERIDICAL because it does not always signal the truth of the two clauses. However we feel it is at least possible to analyse *before* as having two distinct senses, only one of which is veridical. As evidence for this, consider also that with the veridical use of *before*, the contents of the *before*-clause are presupposed, whereas there is no such presupposition in the counterfactual case. Because of this, we choose to omit *before* from the experiment altogether.

4.1.5 The relation type dimension

The third classification experiment involves acquiring the **type** of the relation signalled by a connective. For our purposes, **type** can take three possible values: CAUSAL, TEMPORAL or ADDITIVE. The first two indicate there is always some underlying causal or temporal relation, respectively, whereas ADDITIVE specifies that such a rule need not always be present. These three categories are often present in the literature, although they are often not stated clearly. We will therefore proceed by relating our classes to the literature.

Halliday and Hasan (1976) make a top-level distinction between four major classes of relations: ADDITIVE, ADVERSATIVE, CAUSAL and TEMPORAL. The CAUSAL class includes discourse markers that signal both results and causes, e.g. *so* and *because*, as well as conditional markers, such as the conditional adverbial *then* (as opposed to the temporal adverbial *then*). Halliday and Hasan acknowledge that the cause–result relation is logically different from the antecedent-consequent one, but argue that they are related loosely enough for discourse markers signalling these relations to be "largely interchangeable".

Halliday and Hasan's TEMPORAL class includes markers such as *secondly* and *finally* that signal not temporal relations between events, but sequence relations between textual elements. It also includes markers signalling immediate temporal succession, such as *in a moment*. Such markers are often used when a causal relation exists in the text, as in (4.34).

(4.34) 'Tickets, please!' said the Guard, putting his head in at the window.
In a moment everybody was holding out a ticket. (Halliday and Hasan, 1976, p. 262)

Such markers are classified as TEMPORAL, rather than CAUSAL, presumably because the invariant information they convey is that a temporal relation holds. That is, in other contexts *in a moment* can be used without any causality being implied.

The ADDITIVE class of Halliday and Hasan includes the disjunction *or*, despite a conditional relation being inferable: *A or B* has the implication $\neg A \Rightarrow B$. As we shall soon see, this is in contrast to other analyses in which *or* is explicitly given the attribute CAUSAL (Knott, 1996). Halliday and Hasan's ADVERSATIVE class of markers signal that something is "contrary to expectation". As with *or*, this type of relation can also be analysed as involving an underlying cause (Sanders et al., 1992; Knott, 1996; Kehler, 2002). These differences seem to be a matter of "depth": Halliday and Hasan's classification is based on surface or cohesive relations the writer is signalling to the reader, whereas the other analyses focus more on the logical structures underpinning the relations. This controversy over *or* will lead us to omit it from our experiment.

Martin's (1992) classification has major categories of ADDITIVE, COMPARATIVE, TEMPORAL, and CONSEQUENTIAL, where the last category is closely related to Halliday and Hasan's CAUSAL category. Again, this category includes conditionals, and again *or* is classified as ADDITIVE.

The classification of Sanders et al. (1992) differs significantly from the two just mentioned, in that a separate class of temporal relations is not distinguished. Instead, they claim that temporal relations "belong" to the class of additive relations. Their reasons for this are twofold. Firstly, they claim that the content of clauses, and in particular their tense and aspect, "more or less" determine temporal aspects of meaning. In this respect temporal relations differ from CAUSAL ones, which are not determined by tense and aspect. Temporal order can be reversed without an aspectual shift only if a CAUSAL relation exists, as in (4.35) and (4.36).

(4.35) John has to stand trial. He got a parking ticket. (Sanders et al., 1992, p. 28)

(4.36) John got a parking ticket. He has to stand trial. (Sanders et al., 1992, p. 28)

However it is unclear how Sanders et al. would analyse cases where a connective *reverses* the default temporal ordering, as in (4.37), without any causal implications.

(4.37) A: Did the tension in the boardroom rise before or after the chairman arrived?
B: The tension rose *after* the chairman arrived. (Oberlander and Knott, 1995)

Secondly, for Sanders et al. it is important that the primitives of their theory be "productive", meaning that they can combine freely with other primitives. Crucially for them, temporal rela-

tionships cannot hold between "illocutionary meanings" of segments. As a result, they would disagree with Halliday and Hasan's inclusion of *secondly* as indicating a temporal relation.

Despite this difference with the previously mentioned work, Sanders et al. do include two contrasting primitives of ADDITIVE and CAUSAL, which are closely related to the categories with the same names proposed by Halliday and Hasan. An important difference is that whereas Halliday and Hasan propose a separate ADVERSATIVE category for handling involving contrariness to expectation, for Sanders et al. these arise productively through the combination of the Causal primitive with Negative Polarity. Another similarity with the work described above is that Sanders et al. consider "alternation", as signalled by *or*, to be an Additive relation.

Knott's (1996) classification is similar to Sanders et al.'s in that no separate category of temporal markers is included. Similarly, Knott's INDUCTIVE primitive is closely related to Sanders et al.'s ADDITIVE one. For Knott, discourse relations involve underlying rules, which can be either CAUSAL or INDUCTIVE, the latter involving either some kind of generalisation, or else the failure of such a generalisation to hold. For example, in (4.38) there is a failure to generalise from Bill's liking books to Jill's liking them.

(4.38) Bill and Jill are like chalk and cheese. Bill lives for his books; *whereas* Jill is only interested in Tae Kwan Do. (Knott, 1996, p. 107)

However there is an important difference between Knott's and Sanders et al.'s Causal primitives. Whereas Sanders et al. consider temporal relations to be non-causal, for Knott prototypical signallers of temporal relations such as *meanwhile* and *before* are assigned the attribute CAUSAL, as are markers signalling immediate temporal succession such as *instantly* and *suddenly*. Explanations for this are not given, but a possible reason can be found in other work by the same author. Oberlander and Knott (1995) state that "Whereas *after* can usually have a causal interpretation read into it, this can be defeated by context", following Lascarides and Oberlander (1993). This seems to suggest *after* is more likely to signal a causal relation than a non-causal one, and perhaps similar reasoning was used with other temporal connectives.

Knott's analysis of alternation, e.g. *or*, differs from those of Halliday and Hasan and Sanders et al. Knott classifies *or* as CAUSAL, whereas the others classify it as ADDITIVE. To help understand why Knott does this, consider the example shown in (4.39) containing the related discourse adverbial *otherwise*.

(4.39) Bob put his hands up, *otherwise* Jill would have shot him. (Knott, 1996, p. 117)

Recall that for Knott every discourse relation involves some underlying rule. The crux of Knott's analysis is that in this case the rule is:

$$\neg(\text{Bob puts his hands up}) \rightarrow \text{Jill shoots Bob}$$

This rule is a causal one since it has nothing to do with making generalisations. So for Knott *otherwise* is CAUSAL, and presumably he would have a similar analysis for *or*.

Louwerse's (2001) parameterisation of discourse relations includes a concept of **type**, which can be specified to be either a) CAUSAL, b) TEMPORAL, or c) ADDITIVE. These types of relations indicate either a) a temporal and causal relation, b) only a temporal relation, or c) neither a temporal nor a causal relation, respectively. That is, if there is a causal relation then there must also be a temporal one. According to this analysis, *and* and *but* are ADDITIVE because they do not necessarily imply a temporal relation, while *after* is TEMPORAL because it does not necessarily imply a causal relation. In a manner similar to Halliday and Hasan, the conditional connective *if* is classified as CAUSAL, however in this case there is also the implication that *if* expresses some temporal relation. It is not obvious that this is the case for universals, for example mathematical statements such as (4.40).

(4.40) A triangle has equal sides *if* it has equal angles.

On the other hand, it may be that statements such as this involve some kind of universal temporal quantification, which could make Louwerse's analysis viable. Evidence for this could come from the near-paraphrase using *whenever* shown in (4.41).

(4.41) *Whenever* a triangle has equal angles, it has equal sides.

4.1.6 The relation type gold standard

The gold standard classes of CAUSAL, TEMPORAL and ADDITIVE discourse connectives used in the experiment are shown in Table 4.3. Connectives were classified according to the minimal amount of information they need signal in any situation. For example, although *after* is often used when a CAUSAL discourse relation is intended, minimally it only signals a TEMPORAL relation. The same analysis was also applied to connectives signalling immediate temporal succession, such as *as soon as*, for which there is an even stronger implication of a causal relation (Oberlander and Knott, 1995). We follow the literature in assigning conditionals, such as *if*, to the CAUSAL class.

Several connectives were deliberately omitted from the experiment. These included *or*, due to the disagreement over its type. *While* was omitted because it has two senses, only

ADDITIVE	TEMPORAL	CAUSAL
and, but, whereas	after, as soon as, before, ever since, now, now that, once, until, when, whenever	although, because, even though, for, given that, if, if ever, in case, on condition that, on the assumption that, on the grounds that, provided that, providing that, so, so that, supposing that, though, unless

Table 4.3: Discourse markers in the **type** experiment

one of which involves a temporal relation. *When* is also ambiguous: it can signal temporal simultaneity, temporal succession, and also causation (Moens and Steedman, 1988; Glasbey, 1995). Unlike *while*, however, *when* has been analysed as having a single (underspecified) sense (Moens and Steedman, 1988; Knott, 1996). Moens and Steedman argue that although the meaning of *when* is not primarily temporal, its role is "establishing a temporal focus" (p. 16). For our purposes, the invariant aspect of *when* is to signal a temporal relation of some sort, so it is included in the TEMPORAL class.

4.1.7 The direction of relation dimension

The **type** parameter does not capture the particular temporal ordering signalled by a connective. For example *because* and *so* signal opposite directions of causation, and were not distinguished within the coarse-grained class of CAUSAL connectives. Our concept of **direction** relates to two distinctions made by Halliday and Hasan (1976). They describe the relation between *because* and *so* as involving "Reversal", while *after that* and *before that* are described as SEQUENTIAL and PRECEDING, respectively. In fact, they make a tripartite distinction within the class of simple temporal relations, also including a third subclass SIMULTANEOUS. Martin's (1992) analysis of temporal relations is closely related, except that at the top-level a distinction is made between SIMULTANEOUS and SUCCESSIVE relations, with a finer grained distinction within the latter class determining the temporal order of the succession. However Martin does not make any distinction between the CONSEQUENCE relations signalled by *because* and *so*. This seems to be because Martin is interested in purely in the types of coherence relations that can hold, whereas Halliday and Hasan are more concerned with the elements that can explicitly signal these relations.

Sanders et al. (1992) considers causal relations to have both a BASIC and a NONBASIC form, relating to the textual ordering of the antecedent and the consequent in the text. The relation is BASIC if the antecedent precedes the consequent, otherwise it is NONBASIC. However unlike the other primitives in Sanders et al.'s taxonomy, these cannot be extended to discourse connectives, since a *because*-clause can occur either before or after the cause expressing the consequent, as illustrated by (4.42) and (4.43).

(4.42) *Because* there is a low pressure area over Ireland, the bad weather is coming our way. [BASIC]

(4.43) The bad weather is coming our way *because* there is a low pressure area over Ireland. [NONBASIC]

Nevertheless, despite not being directly applicable to discourse connectives, the distinction they draw is closely related to our parameter of **direction**.

Louwerse (2001) includes a **direction** parameter in his analysis of coherence relations, which can take the values FORWARD or BACKWARD. The distinction here is similar to that made by Sanders et al., in that it refers to the ordering of the text segments.

The concept of DIRECTION that we aim to acquire is most closely related to that of Halliday and Hasan, insomuch as theirs is most closely associated with concrete lexical items, rather than with the types of coherence relations they signal. However we also borrow from Louwerse the idea that a single parameter should underpin the distinction we want to make, by combining freely with other attributes, in particular CAUSAL and TEMPORAL. This desire for productivity takes its roots in the work of Sanders et al., however they do not actually apply it to this particular case, since they consider temporal relations to be a subclass of ADDITIVE ones, and they explicitly exclude the BASIC/NONBASIC parameter from applying to ADDITIVE relations.

4.1.8 The direction of relation gold standard

Our gold standard for the experiment contains two classes, specifying the temporal or causal ordering of the clause introduced by the discourse connective to the other clause. The classes containing *after* and *before* we shall call FORWARD and BACKWARD, respectively. The gold standard classes are shown in Table 4.4. As we discussed in the introduction to the previous experiment, conditionals are often analysed as being closely related to causal connectives. For this reason, our gold standard includes temporal, causal, and also hypothetical connectives.

FORWARD	BACKWARD
after, although, as soon as, assuming that, because, considering that, even if, even though, ever since, for, given that, if, if ever, if only, now, now that, on condition that, on the assumption that, on the grounds that, once, only if, provided that, providing that, seeing as, since, supposing that, the instant, the moment, though, unless	before, in case, in order that, so, so that, then, until

Table 4.4: Discourse markers in the **direction** experiment

The FORWARD class is much bigger than the BACKWARD one, as there seem to be many connectives that introduce conditions, causes, purposes and temporally prior events. In contrast, few connectives seem to introduce consequences and temporally subsequent events. There may be pragmatic reasons for this, for example it might be easier to infer consequences and straightforward temporal succession, so there is less need to signal them explicitly. However there does also seem to be a curious interaction between the **direction** of a causal relation and the syntactic types of discourse markers that that can signal that relation. For example, Knott (1996, p. 178) presents a taxonomy of 20 discourse markers that can introduce cause clauses, and every single one of these is a structural connective. In contrast, of the 45 markers in the taxonomy of markers that can signal results, only 6 are structural connectives (p. 180). Discourse connectives that were deliberately omitted from this gold standard include *while* and *when*, which can both signal temporal overlap.

4.2 Experimental set-up

For each of the semantic dimensions discussed above, classification is performed using the *k* Nearest Neighbour and Naive Bayes classification techniques described in Chapter 2. We use the *k* Nearest Neighbour technique when using lexical co-occurrences as features, because in these cases we can meaningfully represent such co-occurrences using vectors. Before calculating distances between these vectors, we first normalise the vectors so that they represent probability distributions. Note that these do not strictly speaking represent the probability of a word co-occurring with a discourse connective, since in our approach for each discourse connective token we may count co-occurrences with more than one word. Nevertheless, by normalising the frequency counts so that we have a probability distribution, we are able to apply distance

functions that take distributions as their arguments. Essential to this approach is that lexical co-occurrences stand in what is essentially a paradigmatic relation. That is, some words must occur in the clauses linked by a discourse connective, but the choice of actual words may vary.

The situation is somewhat different when we consider the more abstract linguistically motivated features such as tense and negation introduced in Section 3.2.3. It would be meaningless to create a probability distribution that combined measures of, for example, occurrences of the various tense and negation features. Combining tense and negation into a distribution would be like combining chalk and cheese. Instead, tense and negation stand in what is essentially a syntagmatic relation: a clause may take various values for each (e.g. tense can be past or present; a clause is negated or it is not), but the choice for each feature is for the most part independent. In fact, this freedom of combination of the various linguistically motivated features is suggestive of independence. For this reason we apply the Naive Bayes technique when using these features. Naive Bayes is also more suitable in cases when there are fewer features involved, since fewer independence assumptions are required, making Naive Bayes less of an approximation than it would be otherwise. When features are lexical co-occurrences, there are literally thousands of different features, since there are thousand of distinct lexical items. Therefore we do not use Naive Bayes in such cases. An exception arises when we use co-occurrences with other discourse markers as features. In this case, because the set of features is in the order of 350, it is feasible to apply Naive Bayes, although the number of independence assumptions required is still rather high.

The experimental design for each of the machine learning techniques is slightly different too. For k Nearest Neighbour we use a "leave one out" methodology. That is: in turn, for each discourse connective in the gold standard classification, we pretend we do not know its class, and try to accurately reclassify it. We use $k = 1$ for all the experiments, for the following two reasons. Firstly, the class sizes in the experiments are reasonably small, and so the data sets are quite coarse-grained. It therefore makes sense to use a coarser grained classification method too. Secondly, our classes are of unequal sizes, and using a high k in such cases, all else being equal, leads to a bias in favour of the larger classes, at the expense of the smaller. In contrast, when $k = 1$ we can expect a randomly generated data point to be assigned to the various classes with probabilities proportional to the sizes of the classes, giving the correct prior probabilities.

Finally, preliminary analysis of the results revealed two things. Firstly, in general there did not appear to be a substantial difference between the 1NN classifiers using Kullback-Leibler divergence and $Jacc_t$ distance functions, while L_2 appeared to give worse performance. In the interest of conciseness, we therefore only present results using Kullback-Leibler divergence in

this chapter, although some results obtained using L_2 and $Jacc_t$ are provided in Appendix B for the interested reader. Secondly, overall it appeared that verbs and adverbs were the most informative word classes for predicting attributes of connectives (of the classes introduced in Section 3.2). Therefore, when analysing the results of each particular experiment using lexical co-occurrences, we only perform significance tests on the classifiers using a) verbs, b) adverbs, and c) all word classes.

4.3 Experiment 1: Learning polarity

We now present the first classification experiment, in which the goal is to classify discourse connectives according to their **polarity**.

4.3.1 Hypotheses

In Chapter 3 we presented a range of different features of clauses that we suggested could be used to represent aspects of the distributions of discourse connectives. These features ranged from simple lexical co-occurrences, to features representing syntactic, semantic and discourse context. In this experiment we aim to explore the usefulness of these various features for classifying discourse connectives according to their **polarity**. Accordingly, we make the following hypotheses.

Hypothesis 4.1 *Lexical co-occurrences can be used to predict the **polarity** of discourse connectives.*

Hypothesis 4.2 *Co-occurrences of discourse markers can be used to predict the **polarity** of discourse connectives.*

Hypothesis 4.3 *The abstract shallow linguistic features described in Section 3.2.3 can be used to predict the **polarity** of discourse connectives.*

In addition, because negative polarity is related to some deep or underlying negation or inconsistency, we make the following additional hypothesis.

Hypothesis 4.4 *Co-occurrence with negation can be used to predict the **polarity** of discourse connectives.*

	Type of co-occurrences used as features							
Baseline	All POS	VB	RB	AUX	NN	PRP	JJ	IN
0.674	0.721	0.698	0.837	0.605	0.651	0.605	0.721	0.674

Key: VB=main verbs, RB=adverbs, AUX=auxiliary verbs, NN=nouns, PRP=pronouns, JJ=adjectives, IN=prepositions

Table 4.5: Results using the 1NN classifier on lexical co-occurrences. Post-hoc tukey-tests on Baseline, all POS, VB, RB shows no significant differences.

4.3.2 Results

To find support for our hypotheses, we must demonstrate that certain classifiers perform significantly better than some baseline. A repeated measure design was therefore used, whereby we compared the results of applying each classifier to each connective. Our baseline classifier simply assigned each discourse marker to the larger class, i.e. the class with more *types*, in this case POSITIVE POLARITY. The accuracy of this baseline classifier was 0.674.

1NN classifiers The results using the 1NN classifier applied to lexical co-occurrences are shown in Table 4.5. The second column shows the results using co-occurrences with words of all word classes. The remaining columns show the results achieved using co-occurrences with words from just a single class at a time. The best result is achieved using the co-occurrences with adverbs, however despite beating the baseline by over 16%, the post-hoc Tukey test (Howell, 2002) reveals that the difference is not significant.

We next applied a 1NN classifier using co-occurrences of discourse markers. The accuracy of this classifier was 0.814, which is not as good as the best result using co-occurrences with adverbs. However, interestingly, in this case the performance above the baseline is significant ($p < 0.05$). This relates to the fact that we have a repeated measures design, since we attempt to classify each discourse connective with a range of different classifiers. In such cases, the relationship between accuracy and statistical significance is not monotonic; instead the entire confusion matrix comes into play in the calculations, specifically in calculating the amount of error (i.e. unexplained variance). Figure 4.2 shows the confusion matrix for the baseline classifier, and those for the classifiers using adverbs and discourse markers. Crucially, the classifier trained on discourse markers is more similar to the baseline classifier than is the one

	Pos	Neg
Pos	29	0
Neg	14	0

(a) Baseline

	Pos	Neg
Pos	26	3
Neg	4	10

(b) 1NN using adverbs

	Pos	Neg
Pos	28	1
Neg	7	7

(c) 1NN using discourse markers

Figure 4.2: Confusion matrices for three classifiers. Rows indicate the gold standard classes of items; columns indicate the predictions made by the classifiers.

trained on adverbs. This leads to the ANOVA test showing less unexplained variance (i.e. "error") when comparing the baseline and the classifier trained on discourse markers, making it possible for a smaller increase in accuracy to be more significant.

Naive Bayes classifiers One observation that can be made from the results described above is that although the results obtained using co-occurrences with adverbs are not quite statistically significant, it seems that using a subset of lexical co-occurrences can give better results than using co-occurrences with all words. We therefore decided to use Naive Bayes applied to two different sets of discourse markers: either to the entire set, or to a subset that is more likely to be useful. This second set was constructed by calculating which discourse marker co-occurrences had the highest information gain, where information gain was measured using the formula shown in (4.44).

$$InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute) \tag{4.44}$$

Calculations of information gain reported in this chapter were performed automatically using the WEKA machine learning toolkit (Witten and Frank, 2000), and were performed using the entire gold standards. Ideally, it would be preferable to perform feature selection using a held out data set, however this would have been very difficult given the tools available.

The subset of discourse marker co-occurrences with the highest information gains is shown in Table 4.6. The results using Naive Bayes applied to co-occurrences with discourse markers are shown in Table 4.7. Applying the Naive Bayes classifier to just the subset of most informative discourse marker co-occurrences gives an accuracy of over 90%, the best result yet. For comparison, we also show the results using the 1NN classifier (using Kullback-Leibler

Feature	Class with greater mean value	Information gain
though$_{\top}$	POSITIVE	0.306
otherwise$_{\bot}$	NEGATIVE	0.283
but$_{\top}$	POSITIVE	0.276
still$_{\top}$	NEGATIVE	0.262
although$_{\top}$	POSITIVE	0.237
in truth$_{\bot}$	NEGATIVE	0.222
still$_{\bot}$	NEGATIVE	0.212
after that$_{\top}$	NEGATIVE	0.205
in this way$_{\top}$	NEGATIVE	0.205
assuming that$_{\top}$	POSITIVE	0.194
granted that$_{\top}$	NEGATIVE	0.165
in contrast$_{\top}$	NEGATIVE	0.165
by then$_{\bot}$	NEGATIVE	0.165
in the event$_{\bot}$	NEGATIVE	0.165
apart from that$_{\top}$	*equal means*	0.165

Table 4.6: Most informative discourse marker co-occurrences in the subordinate/second clause ($\bot$) and the superordinate/first clause ($\top$)

divergence) on just this subset.

We now turn our attention to using the more abstract linguistic features such as tense and negation. However the results, shown in Table 4.8, achieved using these features did not surpass the results using lexical co-occurrences. A subset of these abstract features giving the highest information gain was also selected, as before. These features are shown in Table 4.9. Using just negation produced slightly better results than using all the abstract features, but none of the results are significantly above the baseline.

Ensembles of classifiers Although the abstract features do not perform well on their own, they might provide useful information in combination with the lexical co-occurrences. In order to explore this, we investigated using ensemble methods for combining the various classifiers, as ensemble methods have been shown to be useful for other machine learning tasks involving nouns (Curran, 2002) and temporal connectives (Lapata and Lascarides, 2004). We adopt

Classifier	Co-occurrences	Result
Naive Bayes	All DMs	0.814
Naive Bayes	Most informative DMs	0.907**
1NN	Most informative DMs	0.698
Baseline		0.674

Table 4.7: Results applying Naive Bayes to discourse markers. Significance measured using a one-way ANOVA: $^{**}p < 0.01$;

Features	Result
All abstract	0.744
Most informative abstract	0.721
Negation in both clauses	0.767
Negation in $\top$ clause only	0.791
Baseline	0.674

Table 4.8: Naive Bayes and abstract linguistic features.

the simplest methodology for combining several classifiers into an ensemble, whereby each classifier is given an equal vote.

One method of constructing successful ensembles is to include individual classifiers which each perform better than chance, and whose errors differ to some degree (Dietterich, 2000). In general, the better performing the individual classifiers, and the more diverse their errors, the better the expected performance of the combined ensemble. It is therefore useful to compare the errors made by the different classifiers reported above. Table 4.10 shows the agreement between the four 1NN classifiers using Kullback-Leibler divergence trained on single POS classes, or on discourse markers, that performed above the baseline. Agreement is measured by the κ statistic (Carletta, 1996), which compares the amount of observed agreement with that expected by chance.

All the correlations are positive, showing greater agreement than by chance, and the highest correlations are between the best two performing classifiers and between the worst two. The high correlation between the classifiers trained on discourse markers and adverbs is perhaps not surprising, since many discourse markers are adverbs. An ensemble of just two classifiers cannot outperform the better of those two, since there is no way to successfully resolve dis-

Feature	Class with greater mean value	Information gain
verbal negation$_\top$	NEGATIVE	0.329
subject negation$_\top$	NEGATIVE	0.270
verb has no args$_\top$	NEGATIVE	0.228
MODALITY=<ABILITY$_\top$,ABILITY$_\bot$>	NEGATIVE	0.194

Table 4.9: Most informative abstract features

		Classifier (accuracy)		
Classifier	(accuracy)	VB (69.8)	RB (83.7)	JJ (72.1)
DM	(81.4)	0.32	0.6	0.49
JJ	(72.1)	0.83	0.54	
RB	(83.7)	0.37		

Table 4.10: Agreement between 1NN classifiers (using the κ statistic, with $N = 43, k = 2$)

agreements between the two. As a result, only ensembles of three and four classifiers were constructed. The results, shown in Table 4.11, show no improvement over the performance of the best classifier in each ensemble. It should be noted that this does not mean the ensembles are worthless. The decision to choose an ensemble rather than any particular classifier requires less prior information about which individual classifier performs best.

Agreement between the different Naive Bayes classifiers, shown in the lower half of Table 4.12, was also always non-negative. However the κ statistic between Naive Bayes and 1NN classifiers was negative in half the cases, and had a maximum value of just 0.16. This shows the two different classification techniques give rise to different errors, which suggests that combining the two types into ensembles may be profitable.

Results from combining different sets of Naive Bayes and 1NN classifiers into ensembles are shown in Table 4.13. Contrary to what one might expect, given the previous discussion, none of the ensembles give as good performance as the 90.7% accuracy achieved earlier.

So far we have only reported the overall accuracy of various classifiers. One factor that this neglects is the performance on individual classes. Table 4.14 compares performance by class of a few well performing classifiers. It shows that better F scores are obtained for the POSITIVE class than for the NEGATIVE one.

Ensemble	Accuracy
JJ + DM + RB + VB	0.779
DM + RB + VB	0.837
JJ + RB + VB	0.721
DM + RB + VB	0.721
DM + RB + JJ	0.837
BASELINE	0.674

Table 4.11: Ensembles of 1NN classifiers

Label	Technique,	Features	(accuracy)	H	G	F	E
A:	1NN,	DM	(81.4)	-0.19	-0.01	-0.14	-0.07
B:	1NN,	JJ	(72.1)	-0.23	-0.14	0.00	0.1
C:	1NN,	RB	(76.7)	-0.15	0.15	0.01	0.16
D:	1NN,	VB	(74.4)	0.02	0.11	-0.14	0.13
E:	Naive Bayes,	all DM	(81.4)	0.02	0.36	0.24	
F:	Naive Bayes,	best DM	(90.7)	0.07	0.00		
G:	Naive Bayes,	best abstract	(72.1)	0.30			
H:	Naive Bayes,	all abstract	(72.1)				

Table 4.12: Agreement with Naive Bayes classifiers (using κ, with $N = 43, k = 2$)

4.3.3 Discussion

Overall, there was mixed support for the hypotheses. Although accuracy rates above the baseline were achieved using lexical co-occurrences, these results were not significant. So we do not have support for Hypothesis 1. However, it should be noted that the gold standard classification is small, containing only 43 discourse connectives. It is more difficult to obtain significant results on such a small set than it would be on a larger one. It is therefore possible such differences may become significant if these experiments were to be re-run on a larger scale.

The best results using co-occurrences with words of a single part of speech were obtained using co-occurrences with adverbs. The high correlation between this classifier and the 1NN classifier trained on co-occurrences with discourse markers suggests that it is likely that it is the

1NN classifiers	Naive Bayes classifiers	Accuracy
DM + RB + VB	abstract	0.837
DM + RB + VB	best abstract	0.826
RB + VB	dms	0.837
RB + VB	best dms	0.860
RB + VB	dms + abstract	0.837
RB + VB	best dms + best abstract	0.860
RB	best dms + best abstract	0.884
VB	best dms + best abstract	0.837
Baseline		0.674

Table 4.13: Ensembles of 1NN classifiers

	Best 1NN classifier (RB)			Best Naive Bayes classifier (most informative DMs)			Best ensemble		
	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*
POSITIVE	0.866	0.897	0.881	0.903	0.966	0.933	0.900	0.931	0.915
NEGATIVE	0.769	0.714	0.741	0.917	0.786	0.846	0.846	0.786	0.815

Table 4.14: Performance on individual classes

S-modifying adverbs that are having the greatest effect. Indeed, many S-modifying adverbs are discourse markers, and can express a wide range of relations in texts. These discourse markers differ from other sentential adverbs in that they require the previous discourse to supply an abstract object for their interpretation (Forbes and Webber, 2002).

Hypothesis 2 was supported by the fact that two classifiers using discourse markers achieved results significantly above the baseline. These were the 1NN classifier using all discourse markers, and the Naive Bayes classifier using just the discourse markers with the highest information gain. Analysis of this set of most informative discourse marker co-occurrences reveals several interesting facts. Firstly, the best predictors of POSITIVE were co-occurrences with *though, but*, and *although*. This shows that if a discourse connective *X* often occurs in constructions with the form (4.45) or (4.47) then *X* is more likely to have the attribute POSITIVE-POLARITY. Examples of such constructions are given in (4.46) or (4.48), respectively

(4.45) Clause1 *but/though/although* (Clause2 *X* Clause3)

Connective	Correctly classified	Connective	Correctly classified
until	1/27	after	27/27
and	1/27	the moment	26/27
but	3/27	so	26/27
because	8/27	insofar as	26/27
since	10/27	considering that	25/27

Table 4.15: Least frequently, and most frequently, correctly classified items using 1NN

(4.46) John was happy, *although* it was raining *and* he had to walk home.

(4.47) Clause1 *but/though/although* (*X* Clause3, Clause2)

(4.48) John left work happy, *although when* he arrived home he wasn't.

This dispreference for having one NEGATIVE POLARITY connective subordinate to another is not dissimilar to the prescription against double negatives.

It is interesting to ask whether certain discourse markers are atypical for their class, in terms of their distributions. In order to answer this, we compared the results of a number of classifiers, to see if some connectives were regularly classified incorrectly. For this error analysis we compared the 27 1NN classifiers that used the different lexical classes and the three different distance functions (Kullback-Leibler divergence, $Jacc_t$ and L_2). Table 4.15 shows the 5 discourse connectives that were correctly classified most frequently, and the 5 that were correctly classified least frequently. Interestingly, *and* and *but*, the two most frequent connectives, were amongst the worst classified. This shows that in terms of their lexical co-occurrences, they behave more like members of the class opposite to which they actually belong. In the case of *and*, it should be remembered that it is often used in cases where the underlying discourse relation does in fact have negative polarity, for example (4.22), repeated below as (4.49). This is what leads Knott to describe *and* as being underspecified with respect to polarity. It is possible, although it would be surprising, that such cases were in the majority in the training data.

(4.49) It was odd. Bob shouted very loudly, *and/but* nobody heard him.

Until was also classified very badly.[1] Although *until* is considered to be NEGATIVE PO-

[1] *Until* and *since* have recently been found to be difficult to distinguish from other temporal connectives on the basis of their co-occurrence distributions (Lapata and Lascarides, 2004).

LARITY by both Knott and Louwerse, the relation it expresses does not obviously contain any underlying negation. In fact, we noted the classification of *until* as NEGATIVE POLARITY was debatable. These empirical results provide further grounds for re-thinking how *until* should be analysed.

Hypothesis 3, that the abstract linguistic features would be useful for the classification task, was not supported. This agrees with previous findings that deeper features do not always give better results (see, for example, Kehler et al., 2004). Even attempts to combine the abstract features with the lexical co-occurrences did not demonstrate any utility of the abstract features. However it is always possible that a more sophisticated application of these features may yield better results.

Finally, we found that of the abstract features, the presence of negation led to the highest information gain, as seen in Table 4.9. However it was only negation in the main clause, or the first of two coordinated clauses, which led to this information gain. That is, in constructions with forms (4.50) and (4.52), exemplified by (4.51) and (4.53), the discourse connective *DC* is more likely to be NEGATIVE POLARITY than if there was no negation.

(4.50) Clause2$_{NEGATED}$ *DC* Clause3

(4.51) John wasn't happy, *but* Sue was.

(4.52) *DC* Clause3, Clause2$_{NEGATED}$

(4.53) *Although* John was happy, Sue wasn't.

This shows that there is a positive correlation between the occurrence of surface negation in a clause, and a discourse relation with NEGATIVE POLARITY taking that clause as an argument. However, despite this correlation, Hypothesis 4 was not supported as the results achieved using negation were not significantly better than the baseline.

4.4 Experiment 2: Learning veridicality

The second experiment concerns learning which discourse connectives are VERIDICAL.

4.4.1 Hypotheses

The hypotheses for this experiment are similar to those for the **polarity** experiment. However this time we hypothesise that modality, rather than negation, will be a useful feature, due to the close relation between **veridicality** and modal status.

	Type of co-occurrences used as features							
Baseline	All POS	VB	RB	AUX	NN	PRP	JJ	IN
0.735	0.857	0.918*	0.673	0.755	0.816	0.796	0.796	0.776

Table 4.16: Results using the 1NN classifier on lexical co-occurrences. Post-hoc Tukey tests did not find any significant differences from the Baseline.

Hypothesis 4.5 *Lexical co-occurrences can be used to predict the **veridicality** of discourse connectives.*

Hypothesis 4.6 *Co-occurrences with other discourse markers can be used to predict the **veridicality** of discourse connectives.*

Hypothesis 4.7 *The abstract shallow linguistic features described in Section 3.2.3 can be used to predict the **veridicality** of discourse connectives.*

Hypothesis 4.8 *Modality in particular can be used to predict the **veridicality** of discourse connectives.*

4.4.2 Results

As for the previous experiment, we evaluate using a repeated measures design of comparisons against a baseline. As before, our baseline is the classifier which assigns each connective to the largest class (in number of types), which in this case is the VERIDICAL class which contains 73.5% of connectives in the experiment.

1NN classifiers The results using the 1NN classifier applied to lexical co-occurrences are shown in Table 4.16. As for the previous task, results above the baseline were achieved using co-occurrences with all parts of speech, however again the difference was not significant. This time it is verbs that gave the best results, with a performance significantly above the baseline. In comparison, adverbs had given the best results for the **polarity** task, whereas here they give results below the baseline. A post-hoc Tukey test applied to the results shown in Table 4.16 found that verbs significantly outperformed auxiliary verbs, pronouns and adverbs, while adverbs also performed less well than prepositions and nouns.

Feature	Class with greater mean value	Information gain
obviously$_{\perp}$	VERIDICAL	0.2303
or$_{\top}$	NON-VERIDICAL	0.2204
now$_{\perp}$	VERIDICAL	0.2139
even$_{\perp}$	VERIDICAL	0.1930
indeed$_{\top}$	VERIDICAL	0.1930
no doubt$_{\top}$	NON-VERIDICAL	0.1930
in turn$_{\top}$	NON-VERIDICAL	0.1717
then$_{\top}$	NON-VERIDICAL	0.1717
once more$_{\top}$	VERIDICAL	0.1687
considering that$_{\top}$	VERIDICAL	0.1687
even after$_{\top}$	VERIDICAL	0.1548
once more$_{\perp}$	VERIDICAL	0.1548
see$_{\perp}$	*classes have equal means*	0.1414
by all means$_{\top}$	NON-VERIDICAL	0.1255
before then$_{\perp}$	NON-VERIDICAL	0.1255
at first sight$_{\top}$	VERIDICAL	0.0765

Table 4.17: Most informative discourse marker co-occurrences in the subordinate/second clause ($\perp$) and the superordinate/first clause ($\top$)

For the **polarity** task, discourse markers had given results significantly above the baseline. For the **veridicality** task they give a performance of 0.796, outperforming the baseline, but not significantly.

Naive Bayes classifiers The application of Naive Bayes to the co-occurrences with discourse markers followed the same line as before. We began by constructing two classifiers. One classifier used co-occurrences with the entire set of discourse markers; the other uses just the subset, shown in Table 4.17, consisting of discourse markers giving the highest information gains. The results are shown in Table 4.18. As in the previous experiment, applying Naive Bayes to just this subset gave significant results that were the best so far for the task. Also as before, applying 1NN to this subset did not give significant results.

We now consider the use of the abstract linguistic features for predicting **veridicality**.

Classifier	Co-occurrences	Result
Naive Bayes	All DMs	0.735
Naive Bayes	Most informative DMs	0.918*
1NN	Most informative DMs	0.776
Baseline		0.735

Table 4.18: Results applying Naive Bayes to discourse markers. Significance measured using a one-way ANOVA: $^*p < 0.01$

Feature	Class with highest mean value	Information gain
verb of main/first clause is *to be*	VERIDICAL	0.284
number of words in main/first clause	VERIDICAL	0.272
number of words in sub/second clause	VERIDICAL	0.272
MODALITY=$< NULL_{\top}, NULL_{\bot} >$	VERIDICAL	0.265
temporal expressions in main/first clause	NON-VERIDICAL	0.183
second person pronoun in main/first clause	NON-VERIDICAL	0.172
second person pronoun in sub/second clause	NON-VERIDICAL	0.172

Table 4.19: Most informative abstract features for each task

Again, we begin by comparing the performance using two sets of features: all the abstract features, and a subset selected according to the principle of information gain, shown in Table 4.19. However as Table 4.20 shows, neither of these were very good predictors of veridicality. It also shows the results using just the nine modality features, in order to test Hypothesis 4.8, as well as just the modality feature with the highest information gain, but neither of these gave much better results.

Ensembles of classifiers It may be that the classifiers described above are better at predicting veridicality in an ensemble than they are individually. As before, we test this by constructing ensembles of classifiers which individually perform above the baseline, but which have differing error patterns. We began by comparing the errors made by the 1NN classifiers that performed above baseline, and the results are shown in Table 4.21. The agreement is almost always non-negative, with nouns (NN), pronouns (PRP) and adjectives (JJ) all showing substantial

Features	Result
All abstract	0.776
Most informative abstract	0.796
Modality only	0.816
Just MODALITY=$< NULL_{\top}, NULL_{\bot} >$	0.776
Baseline	0.735

Table 4.20: Naive Bayes and abstract linguistic features.

Classifier	Classifier (accuracy)					
(accuracy)	VB (91.8)	PRP (79.6)	NN (81.6)	JJ (79.6)	IN (77.6)	DM (79.6)
AUX (75.5)	0.00	0.18	0.10	0.06	0.38	0.30
DM (79.6)	0.19	0.25	0.28	0.25	0.09	
IN (77.6)	0.02	0.21	0.00	-0.15		
JJ (79.6)	0.35	0.50	0.80			
NN (81.6)	0.39	0.54				
PRP (79.6)	0.19					

Table 4.21: Agreement between 1NN classifiers on the **veridicality** task (using κ, with $N = 49, k = 2$)

agreement with each other.

Since verbs give the best performance, we only consider ensembles that include the classifier using verbs. The various ensembles tried are shown in Table 4.22, and include the classifier using verbs in combination with the classifiers with which it has least agreement: those using auxiliary verbs and prepositions. We also tried ensembles of the best performing individual classifiers. However none of the results improved on the performance of the single classifier using verbs.

The agreement between Naive Bayes classifiers and the two best performing 1NN classifiers (using Kullback-Leibler divergence) is shown in Table 4.23. One result that is particularly promising for ensemble creation is that the two most accurate classifiers, namely 1NN using verbs and Naive Bayes using the most informative discourse markers, have a relatively low κ score: less than 0.20. This means that they agree with each other less than 20% more often than would be expected by chance. If we combine these two classifiers with a third one, the role of

Ensemble	Accuracy
VB + AUX + IN + JJ + NN + PRP	0.857
VB + AUX + IN + JJ + NN + PRP + DM	0.857
VB + AUX + IN	0.837
VB + NN + DM	0.857
VB + NN + DM + JJ + PRP	0.857
VB + NN + JJ	0.837
VB + NN + PRP	0.857
BASELINE	0.735

Table 4.22: Ensembles of 1NN classifiers

Label	Technique	Feature	(accuracy)	F	E	D	C	B
A:	1NN	NN	(81.6)	-0.13	0.02	0.05	-0.26	0.39
B:	1NN	VB	(91.8)	0.15	-0.13	**0.18**	-0.14	
C:	NaiveBayes	all dms	(73.5)	0.09	0.04	0.01		
D:	NaiveBayes	best dms	(91.8)	0.00	0.03			
E:	NaiveBayes	best abstract	(79.6)	-0.05				
F:	NaiveBayes	all abstract	(75.5)					

Table 4.23: Agreement with Naive Bayes classifiers (using κ, with $N = 49, k = 2$)

the third will be to try and resolve the cases in which the first two classifiers disagree. Whereas if the two high accuracy classifiers agree with each other, the result of the third classifier is irrelevant.

The result of combining a range of classifiers with the classifiers using 1NN applied to verbs and Naive Bayes applied to the subset of discourse markers is shown in Table 4.24. The most important result here is that the abstract features can be used to boost the performance of an ensemble of classifiers using lexical co-occurrences (including co-occurrences with discourse markers). When we include in the ensemble the Naive Bayes classifier trained on just the modality features, the best result is obtained.

Table 4.25 compares performance by class of a few well performing classifiers. As in the previous experiment, better F scores are obtained for the larger of the two classes, in this case VERIDICAL.

1NN classifiers	Naive Bayes classifiers	Accuracy
VB	best dms	0.918
VB + NN	best dms	0.939
VB	best dms + all abstract	0.980
VB	best dms + best abstract	0.959
VB	best dms + all modality	0.980
VB	best dms + best modality	0.959
Baseline		0.735

Table 4.24: Ensembles of 1NN classifiers

	Best 1NN classifier (VB)			Best Naive Bayes classifier (most informative DMs)			Best ensemble		
	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*
VERIDICAL	0.944	0.944	0.944	1.000	0.889	0.941	0.973	1.000	0.986
NON-VERIDICAL	0.846	0.846	0.846	0.765	1.000	0.867	1.000	0.923	0.960

Table 4.25: Performance on individual classes

4.4.3 Discussion

The results of the **veridicality** experiment differ substantially from what we found in the **polarity** experiment. In the **veridicality** experiment we found that lexical co-occurrences could be used to classify discourse connectives (Hypothesis 4.5), whereas in the **polarity** experiment the result was not significant. The most useful word class for the task was verbs. This is interesting because verbs are syntactically the main predicate of a clause (Merlo and Stevenson, 2001), and as such play a crucial role in the inferring of discourse relations (Kehler, 2002; Asher and Lascarides, 2003).

As for the previous experiment, we compared the errors made by the 27 different 1NN classifiers using lexical co-occurrences (including with DMs), and the best and worst classified connectives are shown in Table 4.26. The connectives most frequently classified correctly are all VERIDICAL, while those least frequently classified correctly are all NON-VERIDICAL. This is not surprising, as most connectives in the gold standard are VERIDICAL, so the nearest neighbour technique has a bias towards assigning connectives to this class. The worst perfor-

Connective	Correctly classified	Connectives	Correctly classified
on the assumption that	6/27	since, as	27/27
if ever	9/27	although, though	27/27
in case	11/27	so, because	27/27
supposing that	12/27	and, but, insofar as	27/27
on condition that	13/27	considering that	27/27
if only	13/27	after, when	27/27

Table 4.26: Least frequently, and most frequently, correctly classified items using 1NN

mances are better than those for the **polarity** experiment, in that in the **polarity** experiment there were two discourse connectives that were classified correctly only once by the 1NN classifiers. Recall from the discussion in Section 4.1.3 that NON-VERIDICAL connectives fall into two major subclasses: conditionals and disjunctions. Given that conditionals far outnumber disjunctions in the gold standard, we might expect the 1NN technique to classify disjunctions poorly. However this is not the case: of the 13 NON-VERIDICAL connectives *or else* is classified correctly most frequently (classified correctly 19 times), while *or* places right in the middle at 7th position (classified correctly 14 times). It may be that close logical relation between conditionals and disjunction ($A \Rightarrow B \equiv \neg A \vee B$) contributes to their having similar co-occurrence distributions.

Hypothesis 4.6 was that co-occurrences with discourse markers could be used to classify discourse connectives. Although none of the classifiers trained on the complete set of discourse markers differed significantly from the baseline, the Naive Bayes classifier trained on the subset of discourse markers with the highest information gain did, achieving 91.8% accuracy. Thus there was support for this hypothesis. Analysis of the discourse markers producing the highest information gains yields some interesting results. The adverbials *obviously* and *indeed*, which intuitively seem to stress the truth of a clause, are associated with VERIDICAL connectives. So too is *once more*; apparently signalling repetition correlates with connectives signalling truth (i.e. veridical contexts). (Note that *once more* presupposes the event occurred before, and veridical contexts are in general more able to resolve presuppositions (Karttunen, 1973).) On the other hand *then* is associated with NON-VERIDICAL, no doubt due to the *if...then* construction, and its variants. *No doubt* and *by all means*, are also associated with NON-VERIDICAL, probably due to the frequency of constructions such as (4.54) and (4.55).

1NN classifiers	Naive Bayes classifiers	Accuracy
VB	best dms + **all modality features**	0.980
VB + **aux**	best dms	0.980

Table 4.27: Ensembles of 1NN classifiers

(4.54) *If* Pompey was in love, *then, no doubt*, so too was his wife.

(4.55) *If* it still looks good to you *then by all means* give it a try.

None of the results achieved using just the abstract features achieved results significantly above the baseline. This is surprising, since marked modality is known to correlate with certain uses of conditionals, e.g. counterfactuals. Thus we have not found support for Hypotheses 4.7 and 4.8. However we did find that ensembles combining classifiers using these abstract features with ones using lexical co-occurrences could boost performance above what was possible using lexical co-occurrences alone. In addition, of these classifiers using abstract features, the biggest improvement was achieved by the one just using the modality features. It appears that the modality features are capturing information that is not captured by the lexical co-occurrences.

Recall from Chapter 3 that the modality features were obtained by analysing the auxiliary verbs in each clause, producing two-dimensional features. However it is possible that a shallower treatment may still extract the crucial information. We therefore constructed one final ensemble, incorporating the 1NN classifier trained on auxiliary verbs. The result, shown in Table 4.27, shows that this classifier appears to have the same boosting effect as the one using modality features. In fact, both the classifiers shown in Table 4.27 only classify one connective incorrectly, and it is the same one: *whenever*. In fact, the universal nature of this connective bears some similarity to certain usages of *if*, as demonstrated by:

(4.56) **Whenever/if** it snows, school is cancelled.

(One difference here is that *whenever* presupposes that it does sometimes snow, whereas *if* does not.) Since no syntactic trees are required to train the classifier using auxiliary verbs as features (just a POS tagger is needed), the ensemble containing it could be argued to be preferable to the one using modality features.

Analysis of which other abstract features yield high information gains shows that the occurrence of second person pronouns in either clause increase the likelihood of a connective being

NON-VERIDICAL. It could be because speakers/writers are more likely to use conditionals when talking/writing about their listeners/readers, as in constructions like (4.57).

(4.57) *If* **you** go down to the woods today, **you**'d better go in disguise!

The information gains also show that if a sentence is longer then it is more likely to be VERIDICAL, although why this should be is not clear. Perhaps NON-VERIDICAL utterances are harder to process, or harder to produce, resulting in a constraint on the amount of other, clause-internal, processing that can be done at the same time. This hypothesis is partially supported by findings that subjects find it easier to comprehend the meaning of the connective *and* than it is to comprehend *or* and *if* (Sacco et al., 2001).

4.5 Experiment 3: Learning relation type

The third experiment aims to learn whether a discourse connective signals an ADDITIVE, TEMPORAL or CAUSAL relation.

4.5.1 Hypotheses

As before, we make four hypotheses concerning the utility of various types of features for the task, which in this case is classifying connectives into the classes CAUSAL, TEMPORAL and ADDITIVE. Hypotheses 4.9 and 4.10 concern the usefulness of simple co-occurrences. Regarding the more abstract features, this time we hypothesise that tense and aspect will be particularly useful features. This is because of the claim that tense and aspect to a large degree determine temporal relations (Sanders et al., 1992).

Hypothesis 4.9 *That lexical co-occurrences can be used to predict the **type** of discourse connectives.*

Hypothesis 4.10 *Co-occurrences with other discourse markers can be used to predict the **type** of discourse connectives.*

Hypothesis 4.11 *The abstract shallow linguistic features described in Section 3.2.3 can be used to predict the **type** of discourse connectives.*

Hypothesis 4.12 *Tense and aspect in particular can be used to predict the **type** of discourse connectives.*

	Type of co-occurrences used as features							
Baseline	All POS	VB	RB	AUX	NN	PRP	JJ	IN
0.581	0.645	0.677	0.742	0.548	0.677	0.613	0.613	0.710

Table 4.28: Results using the 1NN classifier on lexical co-occurrences. Post-hoc Tukey tests did not find any significant differences from the Baseline.

4.5.2 Results

The largest class was the CAUSAL one, and our baseline classifier assigned all connectives to this class, with an accuracy of 58.1%. As with the previous two experiments, we proceed by describing in turn the results achieved using the 1NN, Naive Bayes, and ensemble classifiers.

1NN classifiers The results on the **type** task using the 1NN classifier applied to lexical co-occurrences are shown in Table 4.28. The results are not as high as they were in the previous two experiments, however it must be remembered that the **type** task is inherently harder than the previous two. Firstly, the baseline is lower than in the previous two experiments, as in this case the biggest class (i.e. CAUSAL) does not dominate by as much. Secondly, this experiment involves three classes, CAUSAL, TEMPORAL and ADDITIVE, whereas the previous experiments involved just two.

Although the result was not significant, the best result was achieved using adverbs, as was the case earlier for the **polarity** experiment. Another similarity with the results of the **polarity** experiment is that discourse markers give a significant improvement above the baseline, in this case 0.806 ($p < 0.05$).

Naive Bayes classifiers As for the previous two experiments, Naive Bayes classifiers were constructed using both all discourse marker co-occurrences, and also the subset of most informative co-occurrences, shown in Table 4.30. However, this subset was larger than in the previous experiments, as more co-occurrences gave a high information gain. Of the 54 types of discourse marker co-occurrences selected, 40 (74%) were co-occurrences in the clause immediately following the discourse connective in question. This is more than would be expected by chance (one-tailed sign test, $p < 0.01$).

We report the performance of the classifiers using Naive Bayes and discourse markers in

Classifier	Co-occurrences	Result
Naive Bayes	All DMs	0.581
Naive Bayes	Most informative DMs	0.935**
Naive Bayes	Only most informative DMs in subordinate/second clause	0.935**
Naive Bayes	Only most informative DMs in main/first clause	0.806*
1NN	Most informative DMs	0.581
Baseline		0.581

Table 4.29: Applying Naive Bayes to discourse markers in the **type** experiment. $^{*}p < 0.05$; $^{**}p < 0.01$

Table 4.29. Because there were significantly more informative co-occurrences in the subordinate/second clause, this time we also constructed two classifiers which each used just co-occurrences from each of the two clauses, respectively. The results show that while co-occurrences from both clauses lead to results significantly above the baseline, using co-occurrences from both clauses does not give better results than just using co-occurrences from the subordinate/second clause. Error analysis reveals that in fact these two classifiers make exactly the same predictions.

We now consider the use of the abstract linguistic features for predicting **type**. Again, we begin by comparing the performance using two sets of features: all the abstract features, and a subset selected according to the principle of information gain, shown in Table 4.31. However as Table 4.32 shows, neither of these were very good predictors of **type**. Contrary to our hypothesis that tense and aspect would be useful features (Hypothesis 4.12), the classifier using just these features performed only equal to the baseline. In fact, this classifier assigned 81% of connectives to the CAUSAL class, the same class to which the baseline classifier assigns all markers. This shows that our tense and aspect features do not often predict deviations from the largest class.

Ensemble classifiers Some of the Naive Bayes classifiers already achieve an accuracy of 93.5%, well above the baseline of 58.1%. Of these, the one using just occurrences of discourse markers in the subordinate/second clause as features is the simplest, in that it uses the least features. Therefore, in this section we look at using ensembles of classifiers that include this one, as well as ensembles based simply on lexical co-occurrences.

Feature	Class with greater mean value	Information gain
also$_{\perp}$	ADDITIVE	0.591
again$_{\top}$	CAUSAL	0.486
in addition$_{\perp}$	ADDITIVE	0.480
still$_{\perp}$	ADDITIVE	0.466
altogether$_{\perp}$	CAUSAL	0.457
back$_{\perp}$	CAUSAL	0.449
finally$_{\perp}$	CAUSAL	0.437
only$_{\perp}$	ADDITIVE	0.434
at the same time$_{\perp}$	ADDITIVE	0.433
also$_{\top}$	CAUSAL	0.431
thereby$_{\perp}$	CAUSAL	0.415
back$_{\top}$	TEMPORAL	0.406
once more$_{\top}$	TEMPORAL	0.396
like$_{\top}$	TEMPORAL	0.383
at once$_{\perp}$	CAUSAL	0.382
clearly$_{\perp}$	ADDITIVE	0.382
naturally$_{\perp}$	ADDITIVE	0.382
while$_{\top}$	CAUSAL	0.375
and$_{\top}$	TEMPORAL	0.375
once more$_{\perp}$	TEMPORAL	0.369
clearly$_{\top}$	CAUSAL	0.363
plainly$_{\perp}$	ADDITIVE,CAUSAL	0.363
which was why$_{\top}$	TEMPORAL	0.36
in the first place$_{\perp}$	CAUSAL	0.359
now$_{\perp}$	ADDITIVE	0.347
of course$_{\perp}$	ADDITIVE	0.347
nevertheless$_{\perp}$	ADDITIVE	0.345
admittedly$_{\perp}$	ADDITIVE	0.345
Other co-occurrences in subset of most informative ones: on the one hand$_{\top}$, although$_{\top}$, or$_{\top}$, notably$_{\top}$, by then$_{\perp}$, ultimately$_{\top}$, in contrast$_{\perp}$, unfortunately$_{\perp}$, moreover$_{\perp}$, until then$_{\perp}$, certainly$_{\perp}$, for example$_{\perp}$, in that respect$_{\perp}$, in any case$_{\perp}$, in conclusion$_{\perp}$, apart from that$_{\perp}$, not that$_{\perp}$, anyhow$_{\perp}$, wherein$_{\perp}$, luckily$_{\perp}$, no doubt$_{\perp}$, even then$_{\perp}$, by the same token$_{\perp}$, oh$_{\perp}$, to repeat$_{\perp}$, in spite of this$_{\perp}$		

Table 4.30: Most informative discourse marker co-occurrences in the subordinate/second clause ($\perp$) and the superordinate/first clause ($\top$)

Feature	Class with highest mean value	Information gain
Number of words in main/first clause	ADDITIVE	0.730
Negated subject in subordinate/second clause	CAUSAL	0.671
Number of words in subordinate/second clause	ADDITIVE	0.667
Connective is embedded 7 clauses deep within sentence	TEMPORAL	0.663
Number of clauses embedded beneath subordinate/second clause	ADDITIVE	0.486
MODALITY=<ABILITY$_{\top}$,FUTURE$_{\bot}$>	ADDITIVE	0.486
Verbal negation in subordinate/second clause	CAUSAL	0.434
MODALITY=<ABILITY$_{\top}$,ABILITY$_{\bot}$>	ADDITIVE	0.433
Number of NPs in subordinate/second clause	ADDITIVE	0.433
Negative Polarity Item occurring without negation in subordinate/second clause	CAUSAL	0.396
Negative Polarity Item occurring with negation in subordinate/second clause	CAUSAL	0.371
MODALITY=<FUTURE$_{\top}$,FUTURE$_{\bot}$>	ADDITIVE	0.363
Negative Polarity Items in subordinate/second clause	CAUSAL	0.36
Third person gendered pronouns in main/first clause	TEMPORAL	0.359
MOOD=<DECLARATIVE$_{\top}$,DECLARATIVE$_{\bot}$>	ADDITIVE	0.347
MODALITY=<NULL$_{\top}$,FUTURE$_{\bot}$>	CAUSAL	0.347
MOOD=<INTERROGATIVE$_{\top}$,DECLARATIVE$_{\bot}$>	TEMPORAL	0.329

Table 4.31: Most informative abstract features for each task

Features	Result
All abstract	0.645
Most informative abstract	0.774
Tense and aspect only	0.581
Baseline	0.581

Table 4.32: Naive Bayes and abstract linguistic features.

Classifier	Classifier (accuracy)			
(accuracy)	VB (67.7)	RB (74.2)	NN (67.7)	IN (71.0)
DM (80.6)	0.01	0.08	0.18	0.22
IN (71.0)	0.62	0.11	0.47	
NN (67.7)	0.70	0.22		
RB (74.2)	0.22			

Table 4.33: Agreement between 1NN classifiers on the **type** task (using κ, with $N = 31, k = 2$)

We begin our study of using ensembles to predict **type** by computing the agreement between classifiers using just lexical co-occurrences. Table 4.33 compares just the 1NN classifiers trained on verbs, adverbs, nouns, prepositions and all discourse markers, since these gave the best performances. Overall, agreement is highest between the classifiers trained on verbs and nouns, while the classifiers trained on adverbs and discourse markers have relatively low agreement with all other classifiers. Since the three best performing of these classifiers all show relatively low inter-classifier agreement, they are the obvious candidates for inclusion in an ensemble. However this ensemble does not outperform its best performing member, as shown in Table 4.34. The table also shows that adding additional 1NN classifiers degrades the performance.

The agreement between various Naive Bayes classifiers and the three best performing 1NN classifiers is shown in Table 4.35. The most promising agreement figure is between the the 1NN classifier trained on discourse marker co-occurrences, and the Naive Bayes classifier trained on just the subset of discourse marker co-occurrences in the subordinate/second clause which give the highest information gain. These are the two best performing individual classifiers, yet they have a negative κ agreement statistic, indicating that they agree less than would be expected by chance (for classifiers performing as well as they do). We therefore constructed a range of

Ensemble	Accuracy
DM + RB + IN	0.806
DM + RB + IN + VB	0.774
DM + RB + IN + NN	0.758
DM + RB + IN + NN +VB	0.742
Baseline	0.581

Table 4.34: Ensembles of 1NN classifiers

Label	Technique	Feature	(accuracy)	G	F	E	D	C	B
A:	1NN	dm	(80.6)	0.22	0.32	**-0.09**	-0.2	0.08	0.22
B:	1NN	in	(71.0)	0.06	0.16	-0.10	-0.23	0.11	
C:	1NN	rb	(74.2)	0.11	0.03	0.11	-0.03		
D:	Naive Bayes	all DMs	(58.1)	0.45	0.15	0.17			
E:	Naive Bayes	best DMs in sub/2nd clause	(93.5)	0.29	0.38				
F:	Naive Bayes	best abstract	(77.4)	0.33					
G:	Naive Bayes	all abstract	(77.4)						

Table 4.35: Agreement with Naive Bayes classifiers (using κ, with $N = 31, k = 2$)

ensembles including these two classifiers, but in all cases the accuracy of the ensemble merely matched the accuracy of the better of these two individual classifiers, as shown in Table 4.36.

Table 4.37 compares performance by class on the **type** task of a few well performing classifiers. As before, the worst F scores are obtained for the smallest class.

4.5.3 Discussion

There was significant support for only one of our hypotheses, nevertheless analysis of the results yields some interesting facts. The only lexical co-occurrences that yielded a significant improvement above the baseline were co-occurrences with discourse markers, supporting Hypothesis 4.10. In particular, analysis of information gains showed that it is discourse markers

1NN classifiers	Naive Bayes classifiers	Accuracy
DM + IN	Most informative dms in subordinate/second clause	0.903
DM + RB	Most informative dms in subordinate/second clause	0.903
DM	Most informative dms in subordinate/second clause + all abstract features	0.935
DM	Most informative dms in subordinate/second clause + most informative abstract features	0.935
Baseline		0.581

Table 4.36: Ensembles of classifiers

	Best 1NN classifier (RB)			Best Naive Bayes classifier (most informative DMs)			Best ensemble		
	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*
ADDITIVE	0.250	0.333	0.286	1.000	0.667	0.800	1.000	0.667	0.800
TEMPORAL	0.889	0.800	0.842	1.000	0.900	0.947	1.000	0.900	0.947
CAUSAL	0.778	0.778	0.778	0.900	1.000	0.947	0.900	1.000	0.947

Table 4.37: Performance on individual classes

occurring in the subordinate/second clause of a discourse connective that are most useful for this task. This has important implications for the possibility of extending this task to also include classifying adverbial discourse markers. As discussed in Section 2.2, discourse adverbials take one argument anaphorically, and resolving this anaphor automatically is a difficult task. It would therefore be helpful if discourse adverbials could be classified automatically solely on the basis of the lexical items in the clauses in which the discourse adverbial appears. The results using Naive Bayes classifiers on subsets of discourse marker co-occurrences suggest that this may be possible, although investigating this further is beyond the scope of this thesis.

As in the previous experiments, the performance of the Naive Bayes classifiers using the abstract linguistic features was disappointing. In particular, in contrast to our expectations expressed by Hypothesis 4.12, our tense and aspect features were not useful at all. It is therefore worth reconsidering the relation between the **type** of a discourse connective and the tense and aspect of the clauses it joins more closely. To explore the interaction between **type** and tense,

consider the examples shown in (4.58–4.61). The connectives *because* and *so* both signal a causal relation, albeit with opposite directions of causation. We can see that both can occur with different tenses in the two connected clauses. The general tense schema relating these connectives to the tenses of their connected clauses is shown in (4.62).

(4.58) John went to Edinburgh *because* Sue will be going there.

(4.59) John went to Edinburgh, *so* Sue will be going there.

(4.60) John will be going to Edinburgh *because* Sue went there.

(4.61) John will be going to Edinburgh, *so* Sue went there.

(4.62) {PAST, NON-PAST} *because/so* {PAST, NON-PAST}

In contrast, the temporal connectives *before* and *after* are unable to take different tenses in the clauses they connect, as illustrated by (4.63) and (4.64), resulting in the schemata shown in (4.65).

(4.63) #John went to Edinburgh *before* Sue will be going there.

(4.64) #John will go to Edinburgh *after* Sue went there.

(4.65)
PAST	*before/after*	PAST
NON-PAST	*before/after*	NON-PAST

Recall from Secion 3.2 that our tense features are all two-dimensional, representing the combined tenses of both clauses. We therefore expect that for both *after* and *before* the counts for the features TENSE=<$\text{PAST}_{\top}$,$\text{PRESENT}_{\bot}$> and TENSE=<$\text{PRESENT}_{\top}$,$\text{PAST}_{\bot}$> should be low.

The situation is quite different for aspect. As illustrated by (4.66–4.69), the use of the perfect aspect in a subordinate clause seems determined not by the **type** of the connective, but by the temporal ordering signalled or implied.

(4.66) John dialed the number *after* he had picked up the phone.

(4.67) %John picked up the phone *before* he had dialed the number.

(4.68) John fell *because* he had been pushed.

(4.69) %John was pushed, *so* he had fallen.

Connective	Correctly classified	Connective	Correctly classified
whereas	0/27	providing that	27/27
although	7/27	provided that	27/27
because	7/27	on condition that	26/27
in case	8/27	after	26/27
now	8/27	before	25/27

Table 4.38: Least frequently, and most frequently, correctly classified items using 1NN

In contrast, the use of the perfect aspect in the main clause signals a relation to the preceding text, and is acceptable with all four of these connectives, as shown in (4.70–4.73).

(4.70) John had dialed a number *after* he picked up the phone.

(4.71) John had picked up the phone *before* he dialed the number.

(4.72) John had fallen *because* he was pushed.

(4.73) John had been pushed, *so* he fell.

These examples suggest that although the perfect aspect might occur frequently with particular temporal or causal connectives, it should not correlate with either class as a whole.

The empirical results shown in Table 4.31 show that none of the tense or aspect features led to a large information gain. Instead, it was the modality and mood features that proved most useful. For example, ADDITIVE connectives are more likely to have either ABILITY (as signalled by *can* or *could*) of the future modality in both their clauses, while TEMPORAL connectives are more likely to be used in interrogatives.

The connectives that were classified best and worst by the 27 1NN classifiers experimented with are shown in Table 4.38. The three connectives most frequently classified correctly are all CAUSAL, as would be expected since this is the largest class. However the common temporal connectives *after* and *before* are also handled well. This suggests that in terms of their lexical co-occurrences these connectives can be considered prototypical for their class. The ADDITIVE connective *whereas* was never classified correctly, suggesting that it has quite different co-occurrences from *and* and *but*. Instead, *whereas* was always classified as CAUSAL, showing that its lexical co-occurrences are more similar to CAUSAL connectives.

	Type of co-occurrences used as features							
Baseline	All POS	VB	RB	AUX	NN	PRP	JJ	IN
0.811	0.811	0.676	0.784	0.811	0.730	0.865	0.757	0.757

Table 4.39: Results using the 1NN classifier on lexical co-occurrences

4.6 Experiment 4: Learning direction of relation

The last experiment of the chapter concerns only TEMPORAL and CAUSAL discourse connectives. It aims to learn the direction of the relation being signalled.

4.6.1 Hypotheses

We make four hypotheses concerning the utility of various features for the task, i.e. predicting the **direction** of a connective. The first three hypotheses are the same three that have been used for all four experiments. The final hypothesis is that aspect will be a useful feature for the task. This hypothesis is motivated by the discussion at the end of the previous experiment concerning the interaction between the perfect aspect and temporal and causal ordering.

Hypothesis 4.13 *That lexical co-occurrences can be used to predict the **direction** of discourse connectives.*

Hypothesis 4.14 *Co-occurrences with other discourse markers can be used to predict the **direction** of discourse connectives.*

Hypothesis 4.15 *The abstract shallow linguistic features described in Section 3.2.3 can be used to predict the **direction** of discourse connectives.*

Hypothesis 4.16 *The perfect aspect can be used to predict the **direction** of discourse connectives.*

4.6.2 Results

1NN classifiers Due to the much larger size of the FORWARD class, the baseline performance of 81.1% is much higher than for the previous tasks. Of the 1NN classifiers trained on lexical co-occurrences, shown in Table 4.39, only one beats this baseline, but this is not significant.

The 1NN classifier trained on discourse markers has an accuracy of 0.784, and so does not beat the baseline either.

Naive Bayes classifiers Applying Naive Bayes to co-occurrences with discourse markers gave similar results for this experiment to the previous ones. Table 4.41 shows that a significant result was not achieved using all co-occurrences, but that using only the subset that had the highest information gain enabled a significant result to be achieved. The subset of co-occurrences that were used in this classifier are shown in Table 4.40

The Naive Bayes classifiers using the abstract linguistic features did not perform above the baseline, as can be seen in Table 4.43. The subset of abstract features with the highest information gain is shown in Table 4.42.

Since the best classifier, which applied Naive Bayes to a subset of discourse markers, classified all but one connective correctly, we do not try to improve on this result by constructing ensembles.

Table 4.44 compares performance by class of a few well performing classifiers. As in all the previous experiments, the larger class is handled better than the smaller.

4.6.3 Discussion

The only hypothesis that was supported was that co-occurrences with discourse markers could be used to predict the **direction** of a connective (Hypothesis 4.14). As in the previous experiment, the best result was achieved by training a Naive Bayes classifier on just a subset of the co-occurrences with discourse markers.

Despite the discussion at the end of the previous experiment, our hypothesis that aspect would be a useful feature was not supported. However one factor we did not take into account in that discussion was conditional connectives. Indeed, the perfect aspect has a different interpretation when used in a conditional construction: it is used to signal counterfactuality, as in (4.74).

(4.74) *If* Jane had not come, Pat wouldn't have either.

So it may be that the conditional connectives in the gold standard created so much noise that our predictions were not borne out.

One feature that did turn out to be quite relevant was the depth of embedding of a connective within a clause. Table 4.42 shows that if a connective attaches directly to the main clause of the sentence then it is more likely to be FORWARD, while if it attaches more deeply, e.g. embedded

Feature	Class with greater mean value	Information gain
where$_{\top}$	BACKWARD	0.462
as$_{\top}$	BACKWARD	0.421
once again$_{\perp}$	BACKWARD	0.355
before$_{\top}$	FORWARD	0.355
eventually$_{\perp}$	BACKWARD	0.321
altogether$_{\perp}$	BACKWARD	0.312
beforehand$_{\top}$	BACKWARD	0.308
first$_{\top}$	BACKWARD	0.308
further$_{\perp}$	BACKWARD	0.308
once more$_{\perp}$	BACKWARD	0.308
once$_{\top}$	BACKWARD	0.296
unless$_{\top}$	FORWARD	0.296
while$_{\top}$	BACKWARD	0.296
again$_{\perp}$	BACKWARD	0.272
as long as$_{\top}$	BACKWARD	0.272
in case$_{\top}$	BACKWARD	0.239
or$_{\top}$	BACKWARD	0.231
previously$_{\perp}$	FORWARD	0.231
after$_{\top}$	BACKWARD	0.231
at least$_{\perp}$	BACKWARD	0.220
next time$_{\top}$	BACKWARD	0.220
so that$_{\top}$	BACKWARD	0.220
next$_{\perp}$	BACKWARD	0.220
until$_{\top}$	BACKWARD	0.220
each time$_{\top}$	BACKWARD	0.214
both$_{\top}$	BACKWARD	0.140
even so$_{\perp}$	BACKWARD	0.140
to this end$_{\top}$	BACKWARD	0.140

Table 4.40: Most informative discourse marker co-occurrences in the subordinate/second clause ($\perp$) and the superordinate/first clause ($\top$)

Classifier	Co-occurrences	Result
Naive Bayes	All DMs	0.838
Naive Bayes	Most informative DMs	0.973*
1NN	Most informative DMs	0.892

Table 4.41: Results applying Naive Bayes to discourse markers. Significance measured using F-test: $^*p < 0.05$

Feature	Class with greater mean value	Information gain
Connective is embedded one clause deep	FORWARD	0.251
Connective is embedded two clause deep	BACKWARD	0.251
Connective is embedded three clauses deep	BACKWARD	0.251
Main clause has object but no subject	BACKWARD	0.250
Subordinate clause occurs before main clause	FORWARD	0.231
Subordinate clause occurs after main clause	BACKWARD	0.231
Connective is embedded five clauses deep	BACKWARD	0.231
Subordinate clause contains 3rd person gendered pronouns	BACKWARD	0.231
TENSE=$<\text{NULL}_\top,\text{PRESENT}_\bot>$	BACKWARD	0.220
MOOD=$<\text{NULL}_\top,\text{INTERROGATIVE}_\bot>$	BACKWARD	0.140

Table 4.42: Most informative abstract features. As the gold standard contains no coordinating conjunctions we can speak simply of “subordinate” and “main” clauses.

Features	Result
All abstract	0.811
Most informative abstract	0.757
Perfect aspect only	0.784

Table 4.43: Naive Bayes and abstract linguistic features.

	Best 1NN classifier (PRP)			Best Naive Bayes classifier (most informative DMs)		
	Prec	Rec	*F*	Prec	Rec	*F*
FORWARD	0.931	0.900	0.915	1.000	0.967	0.983
BACKWARD	0.625	0.714	0.667	0.876	1.000	0.933

Table 4.44: Performance on individual classes

two, three or five clauses beneath the topmost S node, then the likelihood of the connective being BACKWARD is increased. Why this should be the case is not clear.

Another feature that was useful was the position of the subordinate clause relative to the main one. (There were no coordinating conjunctions in this experiment.) Specifically, if a subordinate clause occurs before the main clause, the likelihood of the connective being FORWARD is increased. If, however, a subordinate clause occurs after the main clause, the likelihood of the connective being BACKWARD is increased. Example sentences conforming to these trends are shown in (4.75) and (4.76).

(4.75) *After* John picked up the phone, he dialed a number.
(more likely, i.e. $P(\text{FORWARD}|\text{sub clause is preposed}) > P(\text{FORWARD})$)

(4.76) John picked up the phone *before* he dialed a number.
(more likely, i.e. $P(\text{BACKWARD}|\text{sub clause is postposed}) > P(\text{BACKWARD})$)

As these examples illustrate, this trend in the data predicts that clauses are more likely to occur in a certain order, independent of which clause is the complement of a connective. In contrast, (4.77) and (4.78) illustrate choices for the **direction** parameter for which the probability decreases once the position of the subordinate clause wihin the sentence is known.

(4.77) John dialed a number *after* he picked up the phone.
(less likely, i.e. $P(\text{FORWARD}|\text{sub clause is postposed}) < P(\text{FORWARD})$)

(4.78) *Before* John dialed a number, he picked up the phone.
(less likely, i.e. $P(\text{BACKWARD}|\text{sub clause is preposed}) > P(\text{BACKWARD})$)

Furthermore, the more likely order, exemplified by (4.75) and (4.76), has the textual order of the clauses identical to the temporal order of the events. In Sanders et al.'s terms, this shows that when a connective is used the BASIC order of textual segments is preferred to the NONBASIC one.

4.7 Summary

This chapter has examined the automatic acquisition of attributes of discourse connectives. The acquisition task was interpreted within a classification framework, in which the aim was to classify connectives according to the attributes they possess. In all, four different classification tasks were considered, based on independent dimensions of the discourse relations signalled by

Task	Feature	*N*	Accuracy	Kappa (95% confidence interval)
Polarity	Informative DMs	43	0.907	0.780 (0.575–0.985)
Veridicality	*ensemble*	49	0.980	0.949 (0.850–1.048)
Type	Informative DMs	31	0.935	0.877 (0.713–1.014)
Direction	Informative DMs	37	0.973	0.916 (0.755–1.078)

Table 4.45: Best performances on each task

the connectives: **polarity, veridicality, type** and **direction**. These four dimensions re-appear in the literature of discourse connectives, even though the exact definitions of the categories they describe are often disagreed upon, or described imprecisely. To overcome these difficulties, gold standard classes were manually constructed so as to omit controversial or ambiguous discourse connectives.

For each of the four classification tasks, we hypothesised that classification could be performed using various types of features. The main hypotheses were that a) simple lexical co-occurrences, b) co-occurrences with discourse markers, and c) a range of abstract linguistic features would be useful for the classification tasks. The classifiers with the highest accuracy on each task are summarised in Table 4.45. It would be interesting to know which semantic features are easiest to predict using distributional information. However accuracy is not a useful measure of this due to the gold standards being incomparable. Kappa scores can also be used for evaluating classifiers, and they are useful for comparing performances on different tasks because they take into account the level of agreement due to chance. As such, kappa scores are also reported in Table 4.45. Although there appears to be some variation in the kappa scores, we cannot confidently say that any of the differences between the kappa scores are meaningful.

Co-occurrences with discourse markers proved to be the most successful choice of feature. For all four tasks, Naive Bayes classifiers trained on subsets of discourse marker co-occurrences performed well above the baseline. In addition, for the **polarity** and **type** tasks results significantly above the baseline were achieved using 1NN classifiers trained on all discourse marker co-occurrences. For the **veridicality** task, we found that a 1NN classifier applied to co-occurrences with verbs achieved significant results.

The utility of the co-occurrences with other discourse markers raises the question of whether these features might also be useful for discourse processing tasks such as discourse parsing. Unfortunately, a sparseness problem makes these features less useful than might be hoped,

Connective	Frequency	Frequency of co-occurrences with other discourse markers
after	33,010	9,998
and	322,007	104,395
but	166,457	53,080
or	10,141	3,370

Table 4.46: Frequency of discourse marker co-occurrences in the database

because we cannot be guaranteed of having multiple discourse markers within a sentence. Table 4.46 demonstrates this by listing the frequency of discourse marker co-occurrences in our database for a range of connectives. As a consequence, discourse parsers must use other kinds of features for disambiguating discourse marker tokens, at the very least as a backoff option for a model based on discourse marker co-occurrences.

The abstract features did not produce significant results on their own, however it was found that including a classifier trained on modality features could be used to boost performance on the **veridicality** task. The poor performance of the abstract features may be a result of the machine learning technique used, i.e. Naive Bayes. The complete set of abstract features is large and clearly not independent. More sophisticated techniques might be more successful with these features. Nevertheless, the results achieved using shallower features, such as co-occurrences with discourse markers, are encouraging. Complete parse trees are not required for extracting these features (recall that Marcu (1998) identifies discourse markers using finite state techniques), which makes the results more easily transferrable to other languages, for which high performance parsers may not exist.

Various issues have been deliberately ignored for the purposes of our experiments. These include the underspecification of attributes, and also the ambiguity caused by discourse connectives with distinct senses, such as *while*. The guiding principle used in the construction of our gold standard classes was to classify on the basis of information that was invariant throughout all usages of a connective. Obviously, if we had been attempting to classify individual tokens with regards to attributes of the discourse relation they signal, we could not have taken this approach.

Another issue that deserves discussion is the use of the measure of information gain for feature selection. This step was useful for constructing high performance classifiers, and also provided interesting empirical information about the distributions of connectives. For example,

it revealed that surface negation correlates with NEGATIVE POLARITY connectives, and that there is a preference for arranging clauses so that their order mirrors the temporal order of the events they describe. However it must be asked whether using just a subset of features in this way is a valid experimental procedure, or whether it provides unfair assistance. In response to this, firstly note that for each task we followed the same deterministic procedure to select the subsets of features that were then experimented with: we used the attribute selection utility built into the Weka machine learning environment (Witten and Frank, 2000), and used all features which it reported as having a positive information gain. Obviously, if we had tried using many different subsets, perhaps based upon different cutoff values of information gain, then we could not have simply reported the best result and claimed it was valid. Secondly, the use of just a subset of features can be considered an extreme case of weighting features differently, with all weights set to either 0 or 1. Weighting procedures are not uncommon in machine learning (cf. support vector machines, perceptrons) and in our case we made no attempt to optimise the weighting system, or to experiment with different weightings. An obvious alternative choice for weighting would be to weight each feature in proportion to the information gain it provides. We did not do this as the more alternatives that are experimented with, the more difficult it is to claim that the results are significant.

In the next chapter, we consider the task of learning relationships that hold between pairs of discourse connectives. We consider both learning the similarity of pairs of connectives, as well as learning substitutability relationships between connectives.

Chapter 5

Learning relationships between pairs of connectives

The previous chapter concerned the automatic acquisition of attributes of individual discourse connectives. In this chapter we turn our attention to learning relationships between pairs of connectives. As in the previous chapter, we work at the level of types, rather than tokens, as our aim is to acquire information about the lexicon by automatically analysing the distribution of lexical items in a corpus. This chapter directly addresses the main hypothesis of the thesis, which is that discourse connectives with similar meanings also have similar empirical distributions. We demonstrate support for this hypothesis by comparing similarity ratings elicited from human subjects with distributional similarity scores. To further explore the relationship between the distributions of connectives and their meanings, we also consider the more sophisticated task of predicting substitutability relationships between connectives. The concepts of similarity and substitutability are somewhat related: for example, if two words are synonymous then they are both highly similar and (at least under one definition of synonymy) always substitutable for each other. However, as discussed in Chapter 2, there is an inherent asymmetry in the notion of substitution that give rise to four possible (inter-)substitutability relationships. We adopt Knott's (1996) terminology for these four possibilities:

- SYNONYM (x,y): x can always be substituted for y, and vice versa.
- EXCLUSIVE (x,y): x can never be substituted for y, and vice versa.
- CONTINGENTLY SUBSTITUTABLE (x,y): x can sometimes, but not always, be substituted for y, and vice versa.

- HYPONYM (x,y) (or, equivalently, HYPERNYM (y,x)): y can always be substituted for x, but x can only sometimes be substituted for y.

This chapter also introduces a new function of distributions based on the statistical notion of variance, and we provide evidence of its utility in helping to predict substitutability. This novel function constitutes one of the main contributions of the thesis.

In Section 5.1 we present an experiment which elicits human judgements about connective similarity. In Section 5.2 we show that these human judgements correlate positively with distributional similarity. Section 5.3 explores the degree to which distributional similarity can also be used to predict substitutability. Section 5.4 introduces a new variance-based function of probability distributions, and demonstrates that it is sensitive to the substitutability of connectives. Finally in Sections 5.5 and 5.6 we address the task of learning which substitutability relationship holds between a given pair of connectives.

5.1 Experiment 5: Human judgements of connective similarity

5.1.1 Background

The concept of lexical similarity occupies an important role in psychology, artificial intelligence, and computational linguistics. For example, within psychology Miller and Charles (1991) report that:

> [Psychologists] have largely abandoned "synonymy" in favour of "similarity of meaning", "semantic distance", or more generally "semantic similarity". (p. 2)

The same claim is repeated by Charles (2000), suggesting that this trend has continued. Within AI, lexical hierarchies such as WordNet encode semantic similarity through the use of IS-A relations and sets of synonymous, or nearly synonymous, words (which they call "synsets") (Miller, 1990; Fellbaum, 1998). WordNet makes claims about psychological reality, as well as being used in countless NLP applications. Within computational linguistics, work on automatic lexical acquisition is based on the hypothesis that distributional similarity correlates with semantic similarity (Grefenstette, 1994; Curran and Moens, 2002a; Weeds, 2003), a hypothesis that was clearly articulated long before computers were available for performing complicated distributional analysis with large corpora (Rubenstein and Goodenough, 1965; Harris, 1970).

There is ample evidence that subjects can easily rate the similarity of pairs of words such as nouns and verbs. It has also been found that subjects are consistent in their ratings, and

	Similarity ratings				Dissimilarity ratings
Noun pair	R&G	M&C	Resnik	Charles	Charles
gem–jewel	3.94	3.84	3.5	4.00	0.56
food–fruit	2.69	3.08	2.1	2.74	1.34
journey–car	1.55	1.16	0.7	1.70	2.34
coast–hill	1.26	0.87	0.7	1.14	2.80
noon–string	0.04	0.08	0.0	0.95	3.08

Table 5.1: Mean ratings on a scale of 0 to 4 obtained by Rubenstein and Goodenough (1965) (R&G), Miller and Charles (1991) (M&C), Resnik (1999) and Charles (2000)

that there is significant inter-rater agreement. Rubenstein and Goodenough (1965) presented subjects with 65 pairs of nouns such as *noon–string* and *gem–jewel* and elicited semantic similarity judgements on a scale of 0–4. The subjects repeated the experiment two weeks later, and the average correlation of each subject's scores from both sessions was $r = 0.85$. Miller and Charles (1991) elicited similarity judgements for a subset of 30 pairs from Rubenstein and Goodenough's stimuli. The mean scores they obtained had a correlation of 0.97 with the original mean scores. Resnik (1999) elicited judgements for the same 30 pairs, and calculated an inter-rater agreement of 0.90 by using leave-one-out resampling to compare each subject's rating with the mean of those of their peers. Charles (2000) showed that there is a strong negative correlation ($r = -0.97$) between subjects' ratings of semantic similarity and semantic dissimilarity.

Resnik and Diab (2000) performed a similar experiment with 27 verb pairs (e.g. *bathe–kneel*). In this case, two versions of the stimuli were given: one with the verbs given in a sentential context, the other without context. When context was provided, subjects showed a strong tendency to assign lower similarity ratings in general. In both conditions the level of inter-rater agreement was less than that found for nouns: $r = 0.79$ when context was provided; $r = 0.76$ when it wasn't. The difference between conditions may be due to sense disambiguation effects of the contexts. Alternatively, it may even be that subjects rated the semantic similarity of the sentences overall, rather than just the verbs.

The WordNet taxonomy can also be treated as a source of information about noun similarity. Previous studies have proposed a number of similarity metrics based on how nouns are related in WordNet (Hirst and St-Onge, 1998; Leacock and Chodorow, 1998; Resnik, 1995;

Jiang and Conrath, 1997; Lin, 1998b). Budanitsky and Hirst (2001) review these studies, and compare how well their similarity metrics agree with human ratings of similarity, finding that they all compare favourably with an estimated upper bound for the task.

The aim of the experiment presented below is to determine if similar results can be obtained for discourse connectives. That is, do people agree on the degree of semantic similarity of pairs of discourse connectives, such as *despite the fact that* and *even though*?

There are several reasons for believing that judging the similarity of discourse connectives is more difficult than judging the similarity of nouns or verbs. Almost all the nouns used in previous studies refer to concrete objects that people are familiar with, so that people can often identify these objects and even give definitions for the nouns. In contrast, discourse connectives do not have concrete referents, and identifying the relations they signal, let alone defining these relations, can be challenging even for trained linguists. If subjects cannot agree on the semantic similarity of discourse connectives, this would cause problems for our hypothesis that semantically similar connectives are also distributionally similar.

5.1.2 Hypotheses

We have just seen that subjects show high levels of agreement on the semantic similarity of nouns and verbs. Our first hypothesis is that subjects also agree on the similarity of connectives.

Hypothesis 5.1 *Subjects can judge the similarity of pairs of discourse connectives.*

Our next two hypotheses concern the relationship between subjects' similarity ratings and substitutability. We expect that high similarity ratings will be given when two connectives are highly inter-substitutable, and the opposite for non-substitutable pairs of connectives.

Hypothesis 5.2 *Subjects rate pairs of* SYNONYMOUS *connectives as more similar than other pairs of connectives.*

This hypothesis predicts, for example, that pairs such as *but–yet* and *although–even though* should be rated as having high similarity.

Hypothesis 5.3 *Subjects rate pairs of* EXCLUSIVE *connectives as less similar than other pairs of connectives.*

This hypothesis predicts that pairs such as *but–only if* and *although–except when* will be judged to be dissimilar. Note that none of our hypotheses mention HYPONYM or CONTINGENTLY

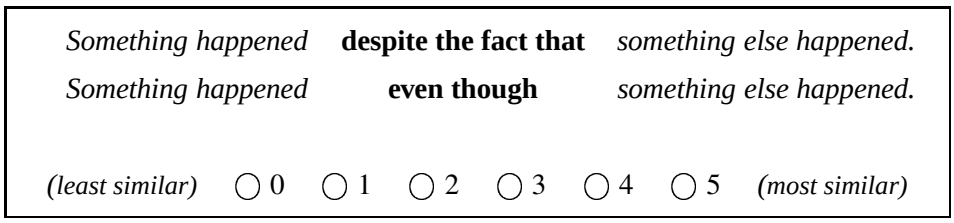


Figure 5.1: An experimental item for eliciting similarity judgements

SUBSTITUTABLE. This is because these relationships both predict partial inter-substitutability, and we do not at this stage make predictions regarding the relative similarity of pairs of connectives in these relationships. (However, later in this chapter we do attempt to model these relationships.)

5.1.3 Methodology

Materials and design From the taxonomy of discourse connectives introduced in Section 3.3, we randomly selected 48 pairs of discourse connectives. The selection was constrained so that there were 12 pairs standing in each of the four substitutability relationships; ambiguous connectives such as *while* were excluded. Each experimental item consisted of the two discourse connectives along with the dummy clauses *Something happened* and *something else happened*. An example stimulus item is shown in Figure 5.1.

The format of the experimental items was intended to balance two conflicting pressures. Firstly, if discourse connectives are presented on their own, without any sentential context, then it may not always be clear how the item can be used to connect clauses. For example, words like *now* and *so* have common uses that are not as discourse connectives, and for a connective like *the moment* it may not be obvious to a naive subject that this can connect clauses at all. Many further examples of confusable discourse connectives were given in Section 3.1. However, if real example sentences are given to illustrate the connective's use, then the subject's judgement may be biased by factors present in those particular example sentences. As a result, the subject may be biased against taking into account the full range of situations in which the connective can be used.[1] For example, if the connective *but* is presented in a context in which it signals

[1] Similar concerns have arisen in eliciting judgements on verb similarity (Resnik and Diab, 2000). The response in that case was to construct two versions of the experimental items: one with example sentences and one without. Budanitsky and Hirst (2001) have pointed out the experimental difficulty in coercing subjects to select a certain sense of a target word without biasing their judgements as to the *a priori* relationship of that word-sense.

the violation of some expectation, the subject might be biased towards this use of *but*, at the expense of other uses such as signalling concessions or semantic opposition.

We opted for a compromise. We present clausal arguments to each connective, to illustrate how it can be used to relate one clause to another. However the semantic contents of the clauses are left grossly underspecified, so that the subject must imagine for themselves what kind of clauses can be connected in this way. This solution is not perfect, since both clauses are always declarative, and the verb *happen* implies the connective relates events rather than states. Nevertheless, it avoids the problems associated with presenting either a bare lexical item on its own or a completely specified context.

Each subject saw each of the 48 pairs of connectives. The items were presented in a different random order for each subject, the ordering of the connectives within each item was also randomised.

Procedure Each participant took part in an experimental session that took approximately 20 minutes. The experiment was conducted remotely over the internet, with subjects accessing the experiment using their web browser. Data obtained over the web has previously been found to give similar results to data obtained in a laboratory (Keller, 2000).

Instructions Before participating in the experiment, subjects were presented with a set of instructions. The instructions began by explaining that there are words and phrases that can connect sentences, and a number of examples of discourse connectives in context were given. Subjects were then told they would be asked to rate the "similarity in meaning" of pairs of connectives. Three example pairs, illustrating high, medium, and low similarity were given. These were *when–while, after–before and because–whereas*, respectively. None of these pairs were also used in the experiment. Subjects were explicitly warned that orthographic similarity should not be taken as implying semantic similarity. The complete instructions, along with all stimulus pairs, can be found in Appendix C.

After the instructions, subjects completed a short questionnaire. Subjects were asked to provide their name, email, age, sex, handedness and the region where they grew up. Subjects were told that if they did not wish to complete the experiment they could submit their partial responses at any time.

Subjects Forty native speakers of English participated in the experiment. Participation was voluntary and unpaid. Of the subjects, 34 were right-handed, 6 left-handed; 15 were female,

Figure 5.2: Correlation of individuals' similarity judgements with means of other subjects

25 male. The age of subjects ranged from 21 to 56; the mean was 36 years.

5.1.4 Results and discussion

One subject completed only 16 of the 48 items. Their ratings are excluded from the correlations of inter-subject agreement, although they are used in the other calculations.

To calculate inter-subject agreement, we used leave-one-out resampling, which is a special case of *n*-fold cross validation. For each subject in turn, we compare their judgement of a pair of connectives with the mean of the judgements of the remaining subjects. Figure 5.2 plots similarity judgements against the mean judgements of other subjects. This technique has previously been used for measuring agreement on judgements of semantic similarity (Resnik, 1999; Resnik and Diab, 2000); other techniques for measuring agreement between multiple subjects are possible, but these would not enable comparisons with findings for nouns and verbs. The average inter-subject correlation was 0.75 (Min = 0.49, Max = 0.86, StdDev = 0.09). These results indicate that subjects agree fairly well on the similarity of pairs of connectives, supporting Hypothesis 5.1. The results are also comparable with inter-subject agreement on verb similarity (Resnik and Diab, 2000); however it is less than the inter-subject agreement on noun similarity (Resnik, 1999). This indicates, as we expected, that rating the similarity of discourse connectives is a more difficult task than rating noun similarity.

In Figure 5.2 a gap can be observed in the mean subjects judgements: none of the exper-

Relationship	Mean	StdDev	Max	Min
SYNONYMY	3.97	1.33	4.82	3.05
HYPONYMY	3.43	1.51	4.56	1.51
CONT SUBS	1.79	1.52	3.10	0.62
EXCLUSIVE	1.08	1.23	2.31	0.55

Table 5.2: Similarity judgements by substitutability relationship

imental stimuli had a mean score between 2.37 and 2.84.[2] That is, the group of subjects as a whole were never evenly split into those who judged a pair to be similar, and those who judged the pair to be dissimilar. Instead, the subjects in effect partitioned the pairs of connectives into two bands, representing high and low similarity. These partitions contain 26 and 22 pairs respectively. Average inter-subject correlation within the high similarity partition was 0.42; within the low similarity partition 0.45. This shows that this partitioning has a major effect on the overall agreement, and that a major part of the agreement can be explained in terms of agreement on whether a pair of connectives has high or low similarity.

The mean similarity ratings for each pair of connectives is given in Appendix C. In Table 5.2 we just give the mean ratings for each of the four substitutability relationships. An Analysis of Variance (ANOVA) was conducted, with the similarity judgements as the dependent variable. The design had repeated measures of each experimental item, with the human subject (Subj) being a between subject variable, and substitutability relationship (Rel) a within subject variable. Main effects were found for Rel ($F(3,44) = 40.057, p < 0.001$) and Subj ($F(38,1672) = 4.767, p < 0.001$), and a crossed effect was found for Subj×Rel ($F(114,1672) = 1.963, p < 0.001$). Post-hoc Tukey tests revealed all differences between relations to be significant (in each case $p < 0.01$), supporting Hypotheses 5.2 and 5.3.

We had not made any hypotheses regarding the relative similarity of pairs of connectives standing in the relationships HYPONYMY and CONTINGENTLY SUBSTITUTABLE. However the results show that HYPONYMY correlates with higher similarity. This can be explained by

[2] Although it has not previously been discussed in the literature, a similar effect can be observed in Rubenstein and Goodenough's (1965) data: there are no mean judgements between 1.82 and 2.37. Similarly, in Miller and Charles's (1991) data there are no mean judgements between 1.66 and 2.82 (although the subset of Rubenstein and Goodenough's (1965) stimuli that they use was randomly selected.) The recurrence of this gap suggests the gap may be an artefact of the experimental design. In particular, it may be due to the requirement that subjects give judgements on an ordinal scale. Magnitude estimation may be a more suitable paradigm for eliciting linguistic judgements (Bard et al., 1996).

Relationship(w_A,w_B)	$A-B$	$A \bigcap B$	$B-A$
SYNONYMY	$\emptyset$	✓	$\emptyset$
HYPERNYM	✓	✓	$\emptyset$
HYPONYM	$\emptyset$	✓	✓
CONT SUBS	✓	✓	✓
EXCLUSIVE	✓	$\emptyset$	✓

Table 5.3: Set theoretic analysis of distributions A and B of words w_A and w_B

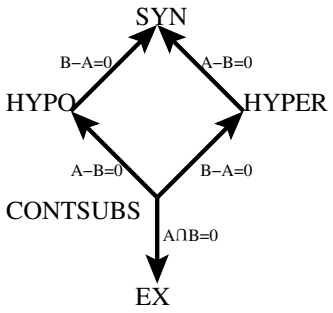


Figure 5.3: Differences between substitutability relationships in terms of empty sets

considering the intersections of the distributions of each connective. Table 5.3 indicates set-theoretic relationships between the sets of contexts A and B for which two connectives w_A and w_B are appropriate. Ticks (✓) indicate that a set is non-empty, while $\emptyset$ indicates an empty set. So, for example, if A and B are EXCLUSIVE, then there is no context in which both are appropriate ($A \bigcap B = \emptyset$), but each can be used in contexts where the other cannot ($A-B \neq \emptyset \neq B-A$). The differences between lines of this Table are represented graphically in Figure 5.3; for example the transition from CONTINGENTLY SUBSTITUTABLE to EXCLUSIVE involves making the change $A \bigcap B = \emptyset$. This Figure implies that CONTINGENTLY SUBSTITUTABLE is more similar to EXCLUSIVE than HYPONYM is, in that to get from the former to EXCLUSIVE just one edge of the graph need be traversed ($A \bigcap B = \emptyset$), whereas to get to EXCLUSIVE from HYPONYM two edges must be traversed. Conversely, HYPONYMY is more similar to SYNONYMY than CONTINGENTLY SUBSTITUTABLE is. This implies an ordering of substitutability relationships: SYNONYMY > HYPONYMY > CONTINGENTLY SUBSTITUTABLE > EXCLUSIVE, and this ordering agrees with the mean ratings per relationship given in Table 5.2.

5.2 Experiment 6: Modelling similarity judgements

5.2.1 Background

Given two words, it has been suggested that the more different their contextual distributions are, then also the more semantically different the words will be (Harris, 1970). Conversely, if two words have the same meaning, then they can be expected to have the same contextual distributions. Cruse (1986) goes even further, arguing that "the meaning of a word is constituted by its contextual relations" (p. 16). However, Weeds (2003) has pointed out that distributional similarity cannot be a sufficient condition for synonymy, since the two most distributionally similar words found by Lin (1998a) were *fall* and *rise*.[3] The following weaker version of the hypothesis has been proposed by Miller and Charles (1991):

Weak Contextual Hypothesis The similarity of the contextual representations of two words contributes to the semantic similarity of those words.

The semantic similarity studies discussed in the previous section have also found evidence that similarity ratings correlate positively with the contextual similarity of the lexical items. However the studies differ in how they measure contextual similarity. On the one hand, Miller and Charles (1991) and Charles (2000) use a measure of discriminability based on a form of sentence completion data. On the other hand, Rubenstein and Goodenough (1965), McDonald (2000) and Resnik and Diab (2000) measure contextual similarity using lexical co-occurrences. Correlation scores are shown in Table 5.4 (Charles (2000) and Miller and Charles (1991) obtain negative correlations because their discriminability measure is greater when items are less similar). In this experiment we aim to determine if the contextual hypothesis also holds for discourse connectives. We will do this by comparing the similarity judgements obtained in the previous experiment with the distributional similarity of connectives.

5.2.2 Hypotheses

One difficulty in extending the results mentioned above to discourse connectives is the question of how to represent the context that a connective appears in. We adopt the approach of Rubenstein and Goodenough, McDonald and Resnik and Diab in using lexical co-occurrences to construct our representations of context. In the previous chapter two features that were found

[3]The main source of data used by Lin was the Wall Street Journal, in which it is likely that these verbs are often used to describe movements in stock prices.

Authors	Word class	Similarity Measure	Correlation
Miller and Charles	Nouns	Discriminability	-0.72
Charles	Nouns	Discriminability	-0.82
McDonald	Nouns	Distributional similarity	0.65
Resnik and Diab	Verbs (with context)	Distributional similarity	0.45
Resnik and Diab	Verbs (without context)	Distributional similarity	0.43

Table 5.4: Contextual correlates of semantic similarity ratings

to give significant results for learning attributes of individual connectives were co-occurrences with verbs and co-occurrences with discourse markers. Specifically, these were occurrences of verbs or discourse markers in the clauses related by a discourse connective, as illustrated by (3.54), repeated below (where the co-occurrences of interest are indicated by bold face).

(3.54) **At first** they might be **offended** *but* **afterwards** they'd **see** I'd **done** them a service.

Here *but* co-occurs with the discourse adverbials *at first* and *afterwards*, and with the verbs *offended*, *see* and *done*.

In this experiment, we hypothesise that these same co-occurrence features can also be used to predict similarity judgements.

Hypothesis 5.4 *There is a linear relationship between semantic similarity ratings obtained from subjects and distributional similarity as measured through co-occurrences with verbs.*

Hypothesis 5.5 *There is a linear relationship between semantic similarity ratings obtained from subjects and distributional similarity as measured through co-occurrences with other discourse markers.*

5.2.3 Methodology

The subjects' similarity judgements from the previous experiment were re-used in this experiment. The lexical co-occurrences were obtained using the method described in Chapter 3. Parse trees were used to obtain syntactic information so that co-occurrences could be indexed by their part of speech and by the clause they occurred in. Co-occurrences were used to calculate distributional similarity, and correlation analysis was used to assess the significance of a linear relationship between distributional similarity and subjects' ratings.

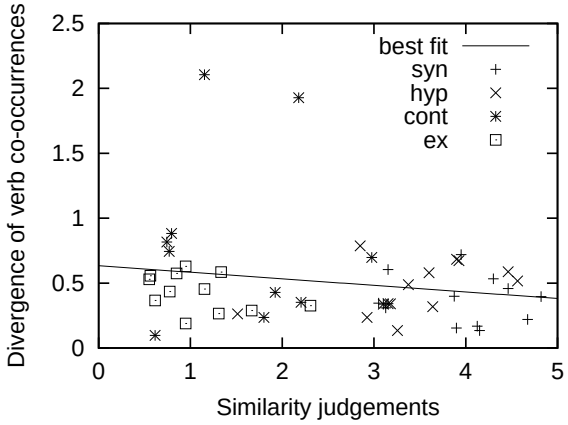


Figure 5.4: Similarity judgements versus KL divergence of co-occurrences with verbs

5.2.4 Results and discussion

A smoothed variant of the Kullback-Leibler divergence function was used to compare distributions (Lee, 2001, with $\alpha = 0.95$). This function is asymmetric, and here we apply it with the arguments ordered according to the alphabetical order of the respective discourse connectives.[4]

The average inter-subject correlation of 0.75 can be considered an upper bound for the task. Figures 5.4 and 5.5 plot the mean similarity judgements against the distributional divergence obtained using co-occurrences with verbs and discourse markers, respectively. Spearman's correlation coefficient for ranked data showed that the correlation is significant when context is represented using discourse markers ($r = -0.52,\ p < 0.001$),[5] but not when context is represented using verbs. (The correlation is negative because KL divergence is lower when distributions are more similar.) Thus Hypothesis 5.5 is supported, but Hypothesis 5.4 is not. For comparison, a third set of distributional representations was also constructed, using co-occurrences with words of all parts of speech. However this model did not produce a significant correlation with the human similarity ratings either. Thus it appears that the co-occurrences of a discourse connective with another discourse marker provides the most information about the semantics of the connective. This is especially so given that there are fewer co-occurrences with other discourse markers than there are co-occurrences with verbs (every clause must have a verb, but need not have a discourse marker).

[4]This is not perfect, but it avoids making arbitrary decisions.

[5]Two outliers can be observed in the graph. The correlation is $r = -0.51$ when these are excluded.

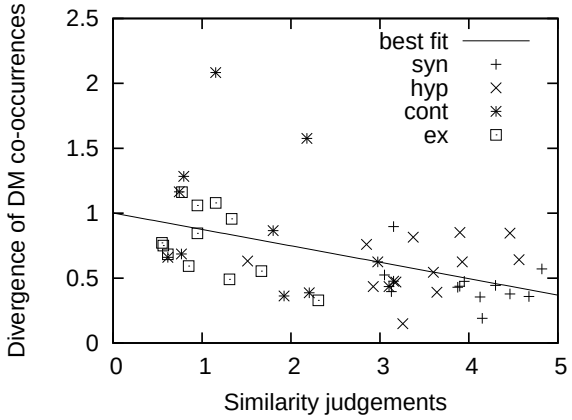


Figure 5.5: Similarity judgements versus KL divergence of co-occurrences with discourse markers

Recall that the human subjects effectively partitioned the pairs of connectives into high and low similarity groups. However the correlation between KL divergence (measured using discourse markers) and human judgements within each of these groups is not significant. This suggests that Kullback-Leibler divergence can be applied to automatically distinguishing between high and low similarity pairs of connectives, but may not be useful for making finer-grained distinctions. This result agrees with previous findings that the similarity of co-occurrence distributions is of little use in distinguishing pairs of nouns of low or moderate similarity (Rubenstein and Goodenough, 1965). However, the small sizes of the subgroups makes it hard to draw any reliable conclusions here.

The magnitude of the correlation between similarity ratings and distributional similarity is slightly higher than comparable results using verbs (Resnik and Diab, 2000). However our achieving this significant correlation relied on choosing a good distributional representation, which in this case meant using co-occurrences with discourse markers. In contrast, the verb study only explored one model of distributional representation, based on labelled syntactic relationships (e.g. "N is the subject of V" or "N is modified by the adjective A"). Exploring further distributional models that take into account more detailed linguistic knowledge might produce higher correlations on either task (cf. Padó and Lapata (2003)).

Many of the theories of discourse coherence discussed in Chapter 2 explicitly group together sets of discourse relations, based upon various principles (for example Grimes, 1975; Longacre, 1983; Halliday and Hasan, 1976; Martin, 1992; Hobbs, 1985). The fact that distribu-

tional similarity correlates with semantic similarity raises the possibility of grouping together similar discourse connectives on empirical principles. To illustrate how this might be done, discourse connectives in the taxonomy introduced in Chapter 3 were clustered automatically using agglomerative hierarchical clustering (Jain et al., 1999). To do this, a symmetric similarity function was defined by applying the Kullback-Leibler divergence function to distributions of co-occurrences with discourse markers, and taking the average of applying with arguments in both possible orders. These scores were then used by the clustering algorithm.

A small selection of the subclusters that were obtained are shown in Figure 5.6, while the entire hierarchy is given in Appendix D. Many of the subclusters are linguistically plausible, for example C69, C32, and C10. Other subclusters are interesting because they seem to ignore certain semantic factors. For example, subcluster C3 of Figure 5.6(k) groups together the causal connectives *so* and *because*, despite their signalling opposite orders of causation (i.e. they take different values on the **direction** dimension discussed in Chapter 4). Similarly, *and* is clustered with several negative polarity connectives in Figure 5.6(g). However despite being underspecified for polarity, *and* does share semantic similarities with these connectives, for example they are all veridical, and none of them indicate a specific temporal relationship. As a result, they can all occur, for example, with a wide range of discourse adverbials signalling different temporal relationships. This may partially explain why *and* has been clustered with them.

In the remainder of this chapter we explore the relationship between the substitutability of pairs of connectives and their empirical distributions.

5.3 Similarity measures for predicting substitutability

The standard technique used in automatic lexical acquisition is to a) calculate the similarity of the distributions of pairs of lexical items, and then b) predict lexical relationships based on this similarity (e.g. Grefenstette, 1994). Many similarity measures have been proposed for this task, including Kullback-Leibler divergence, the cosine metric, confusion probability, Jaccard's coefficient, Jensen-Shannon divergence, Kendall's τ, the L_1 norm (Manhattan distance), the L_2 norm (Euclidean distance) and measures based on the precision and recall of co-occurrences (Weeds and Weir, 2003). Two of these are closely related to lexical substitutability, and so are of particular interest to use in this chapter. They are confusion probability and Kullback-Leibler divergence.

The confusion probability P_C is a formal estimate of "the probability that word w_1' can be

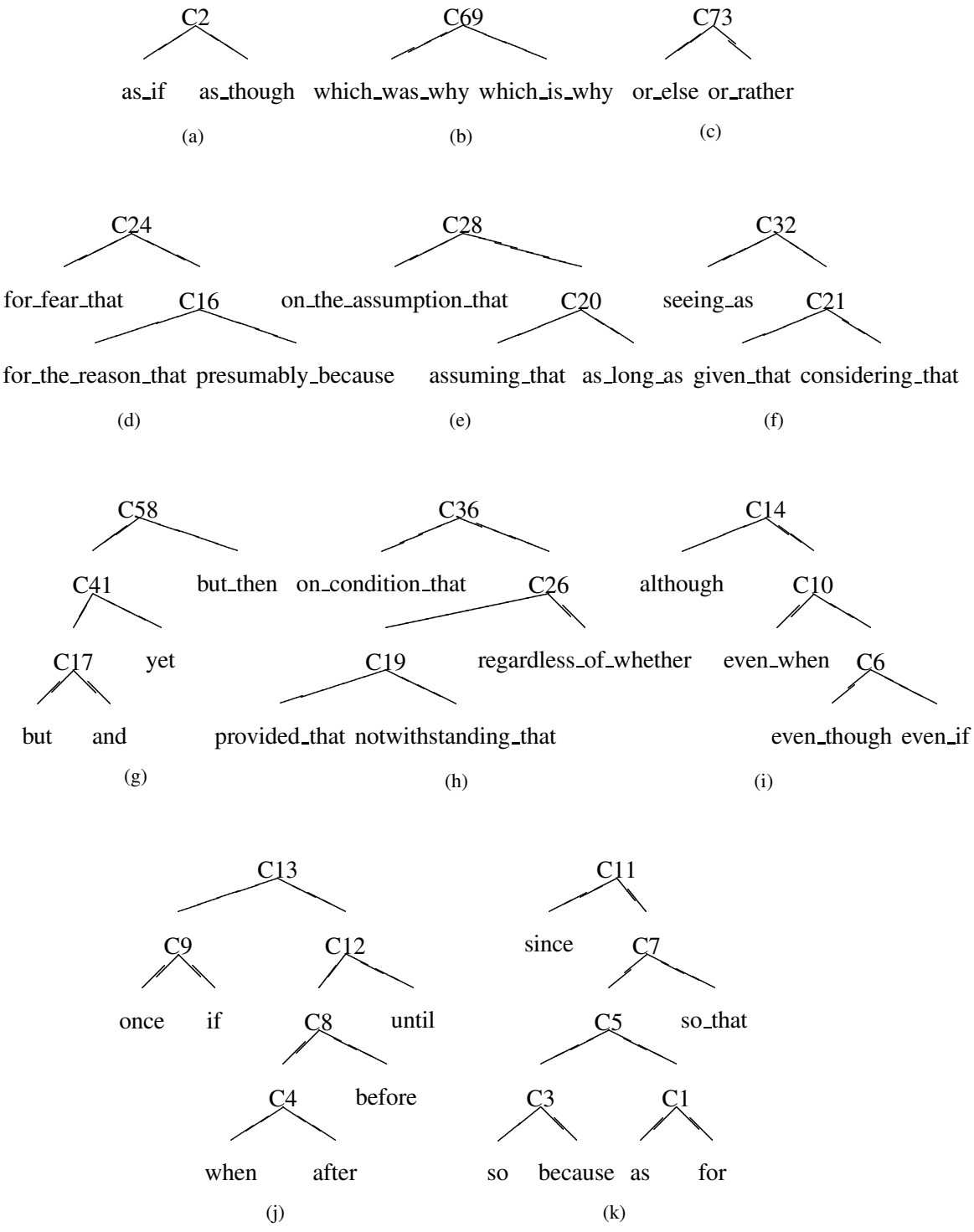


Figure 5.6: Some clusters produced automatically using distributional similarity

substituted for word w_1, in the sense of being found in the same contexts" (Dagan et al., 1999, p. 50), and can be expressed by the equation

$$P_C(w_1'|w_1) = P(w_1') \sum_{w_2} \frac{P(w_2|w_1)P(w_2|w_1')}{P(w_2)} \tag{5.1}$$

Given Dagan et al.'s description, confusion probability would seem ideally suited for the task of predicting substitutability between discourse connectives. However there are several reasons for believing the confusion probability is not the ideal measure for this task, and these reasons are practical, theoretical, and empirical.

Firstly, observe in Equation 5.1 that the calculation of the confusion probability requires estimates of unigram probabilities. In this respect, it is unique among eight similarity measures analysed by Lee (1999). In our case, this causes practical problems due to the difficulties involved in obtaining accurate unigram frequency estimates for discourse connectives. As discussed in Chapter 3, many connectives can easily be confused with other uses of the same words. Our solution was to develop a procedure that identified connectives with a high degree of precision, without being overly concerned with maximising recall. However, as a result the procedure cannot be considered to provide accurate estimates of the relative frequencies of connectives in a corpus. The alternative, to sacrifice precision in order to increase recall, would create more noise in the co-occurrence distributions of the connectives, which would result in inaccurate posterior probabilities in Equation 5.1. Furthermore, our method for sampling example sentences from the web does not straightforwardly facilitate the estimation of unigram probabilities, despite its other advantages that we discussed in Chapter 3.

Secondly, the unigram probability $P(w_1')$ in Equation 5.1 creates an obvious bias towards high frequency words. This bias has been confirmed empirically by Weeds (2003), who found that confusion probability produces neighbours that are 100 to 500 times more likely to have high unigram frequencies than low ones. For predicting symmetric substitutability relationships, such as as SYNONYMY and CONTINGENTLY SUBSTITUTABLE, such an asymmetry in the similarity measure is clearly undesirable. (In contrast, the asymmetry of Kullback-Leibler divergence is not sensitive to unigram probabilities, but just to differences in co-occurrence distributions.)

Thirdly, as a consequence of this reliance on unigram probabilities, the confusion probability has some unusual theoretical properties. One is that the similarity of a word to itself is sensitive to the word's frequency. Another is that a word may not be the most similar word to itself. For example, Dagan et al. show that *fire, role* and *people* are all rated more similar to *guy* than *guy* is to itself.

Fourthly, confusion probability is highly sensitive to the ratio $P(w_2|w_1)/P(w_2)$, whereas most similarity measures are only sensitive to the simpler value $P(w_2|w_1)$. Dagan et al. demonstrate that this can dramatically affect the degree to which different co-occurrences play an important role in determining the similarity value returned. In particular, low frequency words can more easily achieve high values for $P(w_2|w_1)/P(w_2)$, and so play a greater role in deciding similarity. For example, Dagan et al. find that the verbs that give the highest values for $P(w_2|\text{“}guy\text{”})/P(w_2)$ are *electrocute, shortchange* and *bedevil*. It is arguable whether we would want infrequent verbs such as these to have a major effect on calculations of distributional similarity. One thing that is clear, however, is that if low frequency co-occurrences do have a greater impact, then the effects of noise in the frequency counts is exacerbated. In our case, we know that there are several sources of noise in our counts of co-occurrence frequencies. Hence we have another reason to be wary of using the confusion probability.

Finally, confusion probability has not performed favourably in several experiments comparing the performance of a range of similarity measures. Dagan et al. find that Jaccard's coefficient outperforms confusion probability on a word sense disambiguation task. Lee (1999, 2001) conducts pseudodisambiguation experiments and finds confusion probability to be inferior to cosine, Jensen-Shannon divergence, the L_1 norm, Jaccard's coefficient and skewed Kullback-Leibler divergence. Lastly, Weeds (2003) calculates the correlations between WordNet neighbour sets and the predictions of various similarity measures. She finds that confusion probability performs worse than the L_1 norm, Jensen-Shannon divergence, Lee's skewed KL divergence, Jaccard's coefficient, Hindle's (1990) and Lin's (1998a) mutual information based measures, and Weeds' (2003) precision and recall based measures.

Due to this array of arguments against the use of confusion probability, we do not further entertain the idea of using it for our experiments. Instead, we now consider using Kullback-Leibler (KL) divergence, whose definition also relates closely to substitutability. The preceding series of experiments showed that the KL divergence between a pair of connectives correlates with judgements of semantic similarity of that pair, and that these judgements are influenced by the substitutability of the connectives. This gives hope that KL divergence might be used to predict substitutability relationships. Table 5.5 gives statistics on the value of the α-skewed KL divergence for all pairs of connectives in the taxonomy (the arguments to KL divergence were supplied in alphabetical order). Connectives related by SYNONYMY are shown to have significantly less distributional divergence than connectives related by HYPONYMY. These results for discourse connectives relate to similar findings for nouns. Padó and Lapata (2003) find that noun synonyms have somewhat less distributional divergence than superordinate-subordinate

Relationship	N	mean	variance	HYP	CONT.	EXCL
SYNONYMY	20	0.591	0.052	*	*	*
HYPONYMY	52	0.675	0.144		*	*
CONT. SUBS.	878	0.992	0.221			*
EXCLUSIVE	2210	1.001	0.248			

Table 5.5: KL divergences by substitutability, and Tukey test results (* indicates a significant difference)

pairs, although the difference that they found was not significant. The variance column of Table 5.5, however, shows that there is a large degree of overlap in the divergence values for each substitutability relationship. As a result, KL divergence is of limited use in distinguishing between the four substitutability relationships automatically. For example, a KL divergence score of 0.70 is within one standard deviation of the means of each substitutability relationship. Given the greater prior probability of EXCLUSIVE (as exhibited by its greater frequency), a simple Bayesian classifier predicts every pair of connectives to be EXCLUSIVE. In the following section, we propose a new distributional function for helping to predict substitutability.

5.4 Experiment 7: Variation in pointwise entropy

5.4.1 Introduction

Recall from Chapter 2 that KL divergence is formally defined as:

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{5.2}$$

and it measures the "average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite-right distribution q" (Manning and Schütze, 1999, p. 72). Being an information theoretic function, KL divergence also has a natural interpretation in terms of *surprise*. To see this, consider that the definition of $D(p||q)$ can be rewritten as:

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{5.3}$$

$$= E_p(\log \frac{p(x)}{q(x)}) \tag{5.4}$$

$$= E_p(\log \frac{1}{q(x)} - \log \frac{1}{p(x)}) \tag{5.5}$$

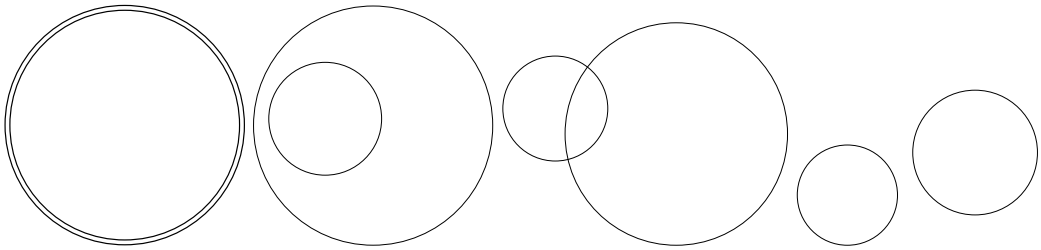
where E_p is the expectation function weighted by the distribution of p. The value $\log \frac{1}{q(x)}$ is known as "pointwise entropy", and can be interpreted as a measure of the surprise in seeing event x, given prior expectations defined by q. So if p and q represent the distributions of two lexical items w_p and w_q, $D(p||q)$ measures how much more surprised we would be, on average, if we saw word w_q in place of word w_p, compared to how surprised we would have been to see w_p there. That is:

$$D(p||q) = E_p(\text{surprise in seeing } w_q - \text{ surprise in seeing } w_p) \tag{5.6}$$

The most commonly used statistical functions are the expectation function (or mean) and the variance function (which measures whether a random variable tends to be consistent or to vary a lot). So as well as measuring the expected pointwise entropy, it is possible that the variance of the pointwise entropy might also be of interest. The variance would provide information about the ways in which two distributions differ (rather than about the degree to which they differ). We now introduce a new function of two probability distributions $V(p,q)$ which measures just this.

$$\begin{aligned} V(p,q) &= Var_p(\text{surprise in seeing } w_q) \\ &= E_p(E_p(\log \frac{1}{q(x)}) - \log \frac{1}{q(x)})^2) \end{aligned}$$

But why should we expect this function to be of interest? Let us now consider how the substitutability of two connectives affects our expectations of the value of V. The substitutability relationships can be represented using the Venn diagrams shown in Figure 5.7. The universe of the Venn diagrams represents the spaces of all discourse contexts in which a discourse connective can be used. We will use these diagrams to illustrate how substitutability relates to our expectations for the value of V.

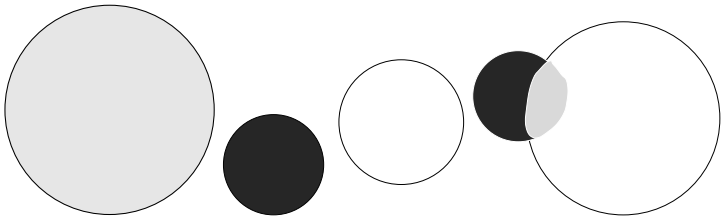


(a) SYNONYMY (b) HYPONYMY (c) CONTINGENTLY SUBSTITUTABLE (d) EXCLUSIVE

Figure 5.7: Venn diagrams representing relationships between distributions

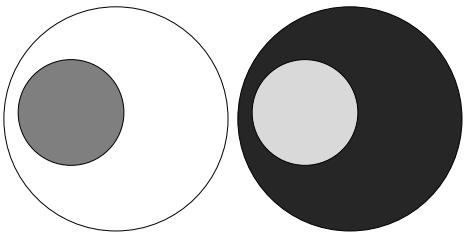
If two connectives are SYNONYMS then each can always be used in place of other. Thus we would always expect a low level of surprise in seeing one connective in place of the other, and this low level of surprise is indicated via light shading in Figure 5.8a. It follows that the variance in surprise is low. On the other hand, if two connectives are EXCLUSIVE then there would always be a high degree of surprise in seeing one in place of the other. This is indicated using dark shading in Figure 5.8b. Only one set is shaded because we need only consider the contexts in which the original connective, rather than the substitutor is appropriate. In this case, the variance in surprise is again low (although the amount of surprise is high). The situation is more interesting when we consider two connectives that are CONTINGENTLY SUBSTITUTABLE. In this case substitutability (and hence surprise) is dependent on the context. This is illustrated using light and dark shading in Figure 5.8c. As a result, the variance in surprise is high. Finally, with HYPONYMY, the variance in surprise depends on whether the original connective was the HYPONYM or the HYPERNYM. Figure 5.9a illustrates that when the HYPERNYM is substituted in the contexts where the HYPONYM is appropriate the variation in surprise is low. However when the HYPONYM is substituted for the HYPERNYM the variance in surprise is high.

From the discussion above, it should be clear that distributional similarity and $V(p,q)$ measure different phenomena. It is worth emphasising this however, as later we will exploit this in order to improve the classification of substitutability relationships. Table 5.6 summarises our expectations regarding the new function V and compares them to expectations for KL divergence. (KL divergence, unlike most similarity functions, is sensitive to the order of arguments



(a) SYNONYMY (b) EXCLUSIVE (c) CONTINGENTLY SUBSTITUTABLE

Figure 5.8: Surprise in substituting connectives



(a) HYPONYMY-1 (b) HYPONYMY-2

Figure 5.9: Surprise in substituting connectives

related by hyponymy (Lee, 1999).) Note that in the case of V these are predictions, based on theoretical expectations, rather than empirical results. In the following sections we will see to what degree empirical support can be found for these predictions.

5.4.2 Hypotheses

Table 5.6 summarises our expectations of typical values for $V(p,q)$ for the different substitutability relationships. These expectations arose from the discussion in above, however in that discussion we implicitly made the simplifying assumption that a connective is equally likely to occur in all discourse contexts in which it is appropriate. As a result, the shading in Figure 5.8 does not vary within regions of the Venn diagrams, whereas in reality there should be variation in expectedness/surprise both within and between regions.

Testing the expectations of Table 5.6 empirically also brings further difficulties, in that

Relationship of w_1 to w_2	Function			
	$D(p\|\|q)$	$D(q\|\|p)$	$V(p,q)$	$V(q,p)$
SYNONYM	Low	Low	Low	Low
HYPONYM	Low	Medium	Low	High
CONTINGENTLY SUBSTITUTABLE	Medium	Medium	High	High
EXCLUSIVE	High	High	Low	Low

Table 5.6: Theoretical expectations for $V(p,q)$ and KL divergence

those expectations are based on abstract discourse contexts, whereas we must rely on approximations of those contexts derived automatically from the presence of concrete linguistic items. The previous experiments in both this chapter and the previous one have shown that distributions of co-occurrences with discourse markers are a useful representation of discourse contexts, and so we will re-use that representation again here. We make the following hypotheses:

Hypothesis 5.6 *$V(p,q)$ returns higher values for* HYPONYMY *and* CONTINGENTLY SUBSTITUTABLE *pairs of connectives than for pairs in the other relationships.*

Hypothesis 5.7 *$V(p,q)$ returns higher values* more consistently *for* CONTINGENTLY SUBSTITUTABLE *pairs of connectives than for pairs in the other relationships.*

Hypothesis 5.8 *When $V(p,q)$ is applied to connectives in an* HYPONYMY *relationship, it is sensitive to the order in which the arguments are applied.*

5.4.3 Methodology

For all pairs of discourse connectives in the taxonomy, we calculated the variance in pointwise entropy ($V(p,q)$), using co-occurrences with discourse markers to represent contextual distributions. In practice, V is not defined if there is an x such that $q(x) = 0$ and $p(x) \neq 0$. To avoid such cases, in all our experiments we use the smoothed variant of V shown in (5.7), inspired by the α-skewed variant of KL divergence (Lee, 1999).

$$V_\alpha(p,q) = V(p, \alpha q + (1-\alpha)p) \tag{5.7}$$

We use the setting $\alpha = 0.95$, however we note that the optimal setting remains an open question.

		f			
Class	N	maximum	minimum	mean	difference-squared
SYNONYM	20	4.44	3.01	3.70	3.29
HYPONYM	52	5.16	2.77	3.96	8.02
CONT. SUBS.	878	4.85	2.53	3.69	7.81
EXCLUSIVE	2210	4.79	2.62	3.70	7.27

Table 5.7: Average values of $f(V(p,q),V(q,p))$, for different functions f

5.4.4 Results and discussion

Because V is asymmetric, we will report the minimum, maximum and mean values of applying V with arguments in both possible orders. In order to estimate sensitivity to the order of the arguments, we also report the value $(V(p,q)-V(q,p))^2$.

As can be seen from Table 5.7, the results only partially support the hypotheses. We had predicted that V would be consistently high when connectives are CONTINGENTLY SUBSTITUTABLE. However this is not supported; in fact for all columns of Table 5.7 the values for CONTINGENTLY SUBSTITUTABLE are not significantly different from those for EXCLUSIVE. HYPONYMY leads to the highest average value for variation in pointwise entropy. The largest differences between $V(p,q)$ and $V(q,p)$ are also for HYPONYMY, as predicted by Hypothesis 5.8. However the differences for HYPONYMY are only significantly greater than those for SYNONYMY ($t = 2.763, p < 0.01$). The "maximum" column shows that the highest values of V are obtained for HYPONYMY and CONTINGENTLY SUBSTITUTABLE. However the difference between CONTINGENTLY SUBSTITUTABLE and EXCLUSIVE is not quite significant ($t = 1.387, p < 0.10$), so Hypothesis 5.6 is only clearly supported for HYPONYMY.

The HYPONYMY relationship appears to produce distinct values of variation in pointwise entropy. It takes higher values in general, and in addition shows more sensitivity to the order of its arguments than SYNONYMY does. This partially supports the expectations that we motivated on theoretical grounds in the above discussion. However, co-occurrences with discourse markers provide only a shallow and imperfect approximation of the discourse context in which a connective appears. It is therefore an encouraging result that they nevertheless provide enough information about the discourse context to partially support our expectations. This provides hope that more sophisticated approximations of discourse context might yield better

support yet.

Earlier in the chapter we saw that KL divergence is significantly correlated with the semantic similarity of pairs of connectives. We also saw that semantic similarity judgements strongly differentiate two sets of substitutability relationships: SYNONYMY and HYPONYMY on the one hand, versus CONTINGENTLY SUBSTITUTABLE and EXCLUSIVE on the other. The results of the current experiment have suggested that $V(p,q)$ may be useful both for distinguishing SYNONYMY from HYPONYMY and for distinguishing the order of the arguments of HYPONYMY. In the following two sections we proceed by applying KL divergence and $V(p,q)$ to the task of learning substitutability relationships.

5.5 Experiment 8: Pseudodisambiguating substitutability relationships

Section 5.1 showed that there is a correspondence between the similarity of a pair of discourse connectives and the type of substitutability relationship that holds between them. In particular, pairs of connectives related by SYNONYMY or HYPONYMY have more similar distributions, and are rated as more similar by subjects, than those that are CONTINGENTLY SUBSTITUTABLE or EXCLUSIVE. Section 5.4 showed that variation in pointwise entropy has a different distribution of values for SYNONYMY and HYPONYMY. This section presents an experiment into using these distributional patterns to distinguish between substitutability relationships. Success on this task can be considered a prerequisite for the automatic acquisition of substitutability relationships.

The four substitutability relationships are distributed very unevenly: by far the most common type of relationship is EXCLUSIVE, while SYNONYMY and HYPONYMY are relatively infrequent. Figure 5.10 shows that there is also a large overlap of the KL divergence scores between the different classes of relationships. In particular, for any range of values of KL divergence, there are more exclusive pairs than other pairs taking those values. It follows that a straightforward classification task would be quite difficult for two reasons. Firstly, the simple baseline of assigning all pairs to the EXCLUSIVE class performs well, with 70% accuracy on the four way classification task. Secondly, the high overlap in KL divergence between classes makes it hard to correctly assign pairs of connectives to the smaller classes without incorrectly assigning many other pairs. As a result, in this section we tackle an easier pseudodisambigua-

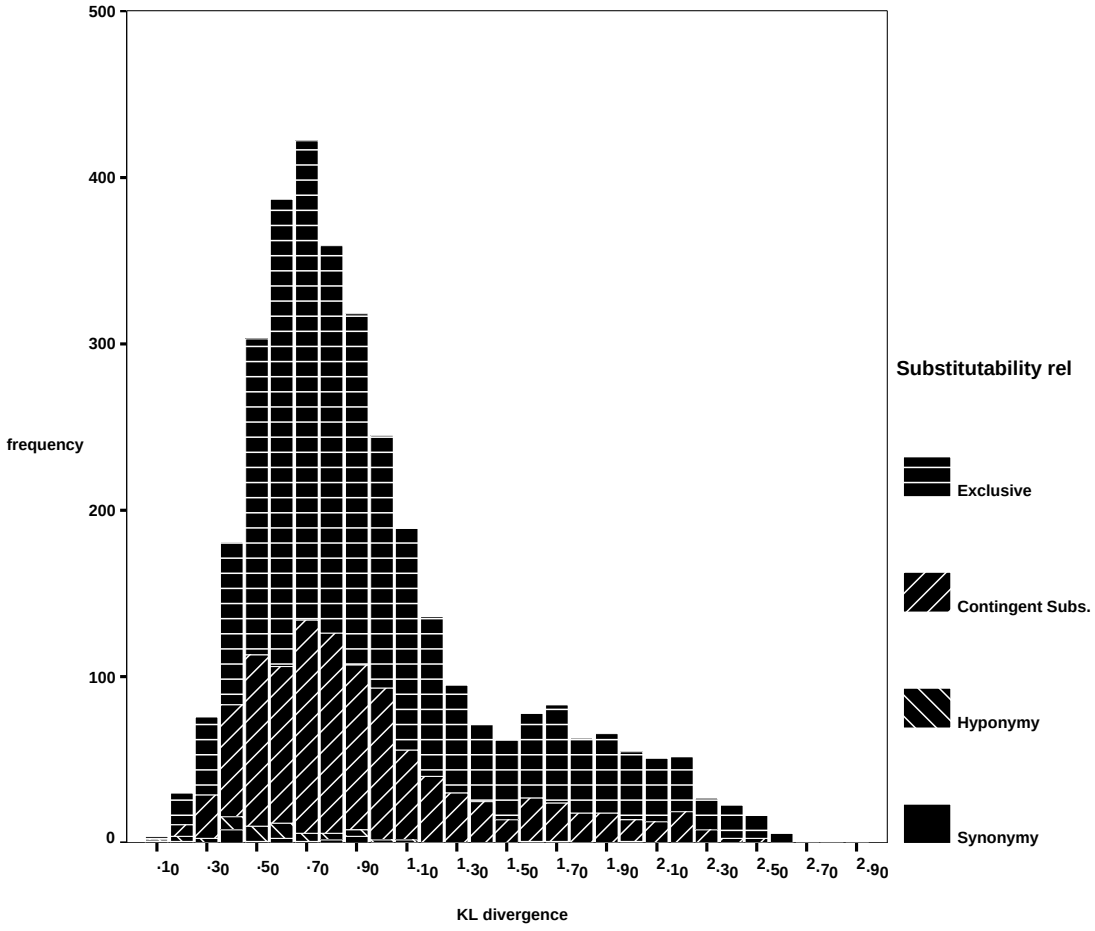


Figure 5.10: Distributions of KL divergences by relation

tion task involving distinguishing two pairs of connectives standing in different substitutability relationships. The more difficult task of predicting which relationship holds for any individual pair is attempted in the next chapter.

5.5.1 The task

Two pseudodisambiguation tasks were attempted, each involving distinguishing substitutability relationships using distributional information. The first task involved distinguishing between SYNONYMY and HYPONYMY. Given three discourse connectives p, q, q', such that SYN-

ONYM(p, q) and either HYPONYM(p, q') or HYPONYM(q', p), the task was to decide which of q and q' was the SYNONYM of p. For example, given that *although* is a SYNONYM of one of *even though* and *notwithstanding that*, and has the other as a HYPONYM, the task is to decide between the following alternatives:

a) SYNONYM (although,even though) and HYP(although,notwithstanding that), or

b) SYNONYM (although,notwithstanding that) and HYP(although,even though).

In this case it is (b) that is the correct answer.

The second task was identical in nature to the first, however here the relationship between p and q was either SYNONYMY or HYPONYMY, while p and q' were either CONTINGENTLY SUBSTITUTABLE or EXCLUSIVE. In combination, the two tasks are equivalent to predicting SYNONYMY or HYPONYMY from the set of all four relationships, by first distinguishing these from CONTINGENTLY SUBSTITUTABLE and EXCLUSIVE, and then making a finer-grained distinction between the SYNONYMY and HYPONYMY. These tasks allow us to explore methods for distinguishing substitutability relationships on the basis of distributional features, without tackling the more difficult task of actually predicting substitutability.

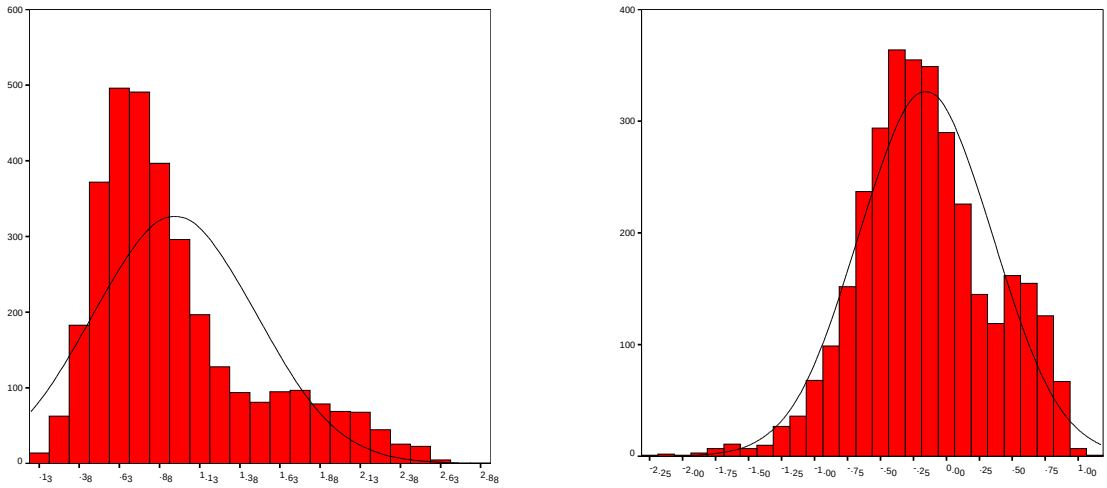
5.5.2 Materials

The taxonomy of discourse connectives introduced in Section 3.3 was used to extract evaluation data for the task. There were 46 triples used in the first task, and 10,912 triples in the second task. The reason for the large difference is that there were relatively few triples of connectives satisfying the requirements of the first task.

5.5.3 Method

Lexical co-occurrence data were used to calculate $D(p||q)$ and $V(p,q)$. Smoothing was used to prevent problematic zero denominators, as discussed in Section 5.4.3. These functions are both asymmetric, so we obtained symmetric functions by taking the a) maximum, b) minimum, and c) square of the difference of applying the function with arguments in both possible orders.

Our methods for the pseudodisambiguation tasks rely on modelling the distributions of values of $D(p||q)$ and $V(p,q)$. As such, preliminary analysis of the distribution of values of $D(p||q)$ and $V(p,q)$ was carried out. Figure 5.11a plots the frequency of values of the KL divergence function applied to all pairs of connectives in the taxonomy, and shows that the



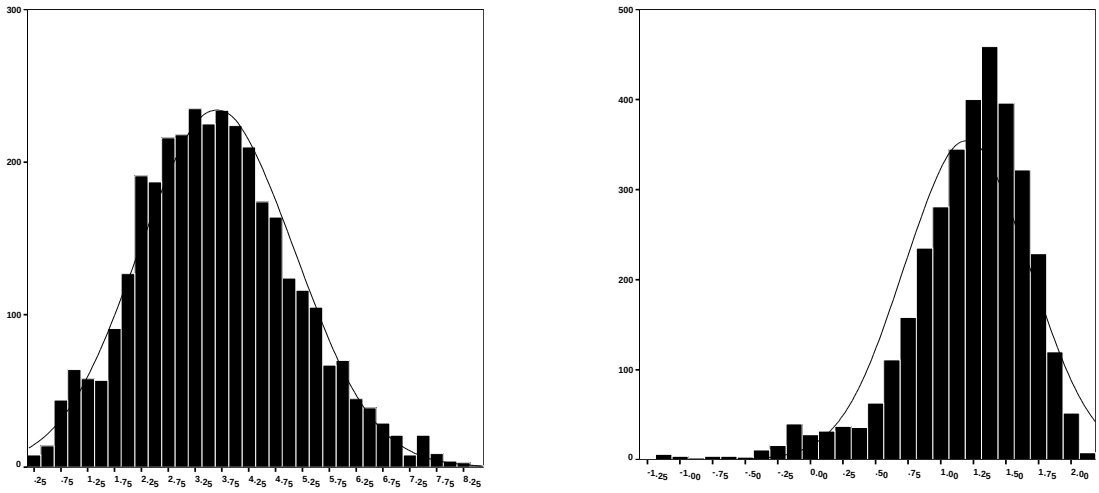
(a) $D(p||q)$ (b) $\log D(p||q)$

Figure 5.11: Fitting normal curves to KL divergence

best-fitting normal curve approximates the data poorly. Figure 5.11b shows that a normal curve fits the data of $\log D(p||q)$ much better, i.e. the $D(p||q)$ data is approximately distributed log-normally. This is not surprising, given that KL divergence is right-skewed due to its having a lower bound of zero. In contrast, variation in pointwise entropy is approximated well by a normal curve, as can be seen by comparing Figures 5.12a and 5.12b.

The distributions of values of $D(p||q)$ and $V(p,q)$ were used to construct two fitness functions. These functions each take four arguments: a function of two distributions f, a pair of connectives d and d' and a substitutability relationship *rel*. The fitness functions return a numerical value which relates to how typical the distributional data of the two given connectives is for pairs of connectives in the given relationship.

Linear fitness function The simpler of the two fitness functions simply compares the value $f(d,d')$ to the mean value μ_{rel} obtained by applying f to all pairs of connectives in the relationship *rel*. The Euclidean distance function is used to compare $f(d,d')$ to the mean:



(a) $V(p,q)$ (b) $\log V(p,q)$

Figure 5.12: Fitting normal curves to variation in pointwise entropy

$$Linear(f,d,d',rel) = |f(p_d,p_{d'}) - \mu_{rel}| \tag{5.8}$$

where p_x is the probability distribution of connective x, and where

$$\mu_{rel} = E_{rel(x,y)}(\{f(d_x,d_y)\}) \tag{5.9}$$

The value of the linear fitness function is smaller if the distributions of the connectives are related in a manner typical for that relationship. Otherwise it is larger. This fitness function was applied to the first pseudodisambiguation task as follows. Suppose the three connectives are d,d' and d''. Then we make the following predictions:

1. SYNONYM (d,d') and HYPONYM (d,d'') if

$$\frac{Linear(f,d,d',\text{SYNONYM}) + Linear(f,d,d'',\text{HYPONYM})}{Linear(f,d,d',\text{HYPONYM}) + Linear(f,d,d'',\text{SYNONYM})} < 1$$

2. HYPONYM (d, d') and SYNONYM (d, d'') if

$$\frac{Linear(f, d, d', \text{SYNONYM}) + Linear(f, d, d'', \text{HYPONYM})}{Linear(f, d, d', \text{HYPONYM}) + Linear(f, d, d'', \text{SYNONYM})} > 1$$

Gaussian fitness function This fitness function takes both the mean and the variance of the values of f into account, by using Gaussian models of the data.

$$\begin{aligned} Gauss(f, d, d', rel) &= n(f(p_d, p_{d'}); \mu_{rel}, \sigma_{rel}) \\ &= \frac{1}{\sigma_{rel}\sqrt{2\pi}} e^{-(f(p_d, p_{d'}) - \mu_{rel})^2 / 2\sigma_{rel}^2} \end{aligned}$$

where again p_x is the probability distribution of connective x, and where:

$$\begin{aligned} \mu_{rel} &= E_{rel(x,y)}(\{f(d_x, d_y)\}) \\ \sigma_{rel} &= \sqrt{Var_{rel(x,y)}(\{f(d_x, d_y)\})} \end{aligned}$$

In contrast to the linear fitness function, the value of the Gaussian fitness function is larger if the distributions of the connectives are related in a manner typical for that relationship. Otherwise it is smaller. This fitness function was applied to the first pseudodisambiguation task as follows. Suppose the three connectives are d, d' and d''. Then we make the following predictions:

1. SYNONYM (d, d') and HYPONYM (d, d'') if

$$\frac{Gauss(f, d, d', \text{SYNONYM}) Gauss(f, d, d'', \text{HYPONYM})}{Gauss(f, d, d'', \text{HYPONYM}) Gauss(f, d, d', \text{SYNONYM})} > 1$$

2. HYPONYM (d, d') and SYNONYM (d, d'') if

$$\frac{Gauss(f, d, d', \text{SYNONYM}) Gauss(f, d, d'', \text{HYPONYM})}{Gauss(f, d, d'', \text{HYPONYM}) Gauss(f, d, d', \text{SYNONYM})} < 1$$

If we consider that the ratio of two values of a Gaussian function is a likelihood ratio, then it is clear that the fractions in the above inequalities also correspond to likelihood ratios.

Function	Features	Symmetrisation	Definition
D	DMs	Max	$max(D(p\|\|q), D(q\|\|p))$, where p and q represent distributions of co-occurrences with discourse markers
V	all words	Diff^2	$(V(p,q) - V(q,p))^2$, where p and q represent distributions of co-occurrences with all word types

Table 5.8: Example parameter settings for f

These two fitness functions were also adapted to the second task, so that in the inequalities above we replace SYNONYM with "SYNONYM OR HYPONYM", and we replace HYPONYM with "CONTINGENTLY SUBSTITUTABLE OR EXCLUSIVE". A range of different functions f were experimented with, varying along three independent parameters. The first parameter specified the co-occurrence features that were used. The two possibilities tried were a) co-occurrences with discourse markers, and b) co-occurrences with all words in the related clauses. Other choices are of course possible, but these choices cover both a naive approach (co-occurrences with all words) and a more sophisticated approach (co-occurrences with discourse markers). The second parameter specified the basic function of two distributions. This was either KL divergence ($D(p||q)$) or variation in pointwise entropy ($V(p,q)$). The third specified the method of producing a symmetric function from the asymmetric D and V. The three possibilities tried here were a) Minimum, b) Maximum, and c) the square of the differences. Some example parameters settings for f are given in Table 5.8, along with the value that they represent in full.

5.5.4 Results and discussion

Leave-one-out cross validation was used. For each triple $\langle p, q, q' \rangle$, the data concerning the pairs p, q and p, q' were held back, and the remaining data used to construct the models. The results using the Linear fitness function are shown in Table 5.9; those using the Gaussian fitness function in Table 5.10. Higher levels of accuracy are achieved on the task distinguishing SYNONYMY from HYPONYMY, however there is not a single function that performs best on both tasks. If we compare the performance of $D(p||q)$ on the two tasks, we see that using co-occurrences with all words gives better performance distinguishing SYNONYMY from HY-

Function	Features	Symmetrisation	SYNONYMY VS HYPONYMY	SYN/HYP VS EX/CONT. SUBS.
$D(p\|\|q)$	DMs	Max	0.717	0.766
$D(p\|\|q)$	DMs	Min	0.717	**0.767**
$D(p\|\|q)$	all words	Max	**0.891**	0.719
$D(p\|\|q)$	all words	Min	0.870	0.732
$V(p,q)$	DMs	Max	0.848	0.548
$V(p,q)$	DMs	Min	0.630	0.646
$V(p,q)$	DMs	Diff^2	0.717	0.520
$V(p,q)$	all words	Max	0.717	0.520
$V(p,q)$	all words	Min	0.522	0.555
$V(p,q)$	all words	Diff^2	0.804	0.513
Baseline			0.500	0.500

Table 5.9: Pseudodisambiguation using Linear fitness functions

Function	Features	Symmetrisation	SYNONYMY VS HYPONYMY	SYN/HYP VS EX/CONT. SUBS.
$D(p\|\|q)$	DMs	Max	0.500	0.760
$D(p\|\|q)$	DMs	Min	0.413	**0.761**
$D(p\|\|q)$	all words	Max	0.804	0.712
$D(p\|\|q)$	all words	Min	**0.848**	0.732
$V(p,q)$	DMs	Max	**0.848**	0.606
$V(p,q)$	DMs	Min	0.587	0.508
$V(p,q)$	DMs	Diff^2	0.717	0.551
$V(p,q)$	all words	Max	0.609	0.512
$V(p,q)$	all words	Min	0.478	0.521
$V(p,q)$	all words	Diff^2	0.783	0.557
Baseline			0.500	0.500

Table 5.10: Pseudodisambiguation using Gaussian fitness functions

Function 1	Function 2	SYNONYMY VS HYPONYMY	SYN/HYP VS EX/CONTSUBS
$D(p\|\|q)$,DMs,Max	$V(p,q)$,DMs,Max	0.761	0.762
$D(p\|\|q)$,all words,Max	$V(p,q)$,DMs,Max	0.891	0.728

Table 5.11: Pseudodisambiguation by combining Gaussian functions

PONYMY, but using co-occurrences with discourse markers gives better performance distinguishing "SYNONYM OR HYPONYM" from "CONTINGENTLY SUBSTITUTABLE OR EXCLUSIVE". This suggests that the discourse context (as indicated by other discourse markers) is better at discriminating high similarity pairs of connectives from low similarity ones, but that the semantic contents of the clauses (as indicated by all word co-occurrences) are more useful for making the finer-grained distinction between SYNONYMY and HYPONYMY.

Variation in pointwise entropy is useful for discriminating SYNONYMY and HYPONYMY. The best classifier based on it had an accuracy of 0.848, which is not significantly different from the best classifiers based on KL divergence. However variation in pointwise entropy did not perform so well on the coarser task of distinguishing SYN/HYP from EX/CONT. SUBS.

The differences in results when using the Linear fitness function and the Gaussian fitness function are not great, and the same general trends in results can be seen. However, as mentioned above, one advantage of using the Gaussian fitness function is that ratios of values of Gaussian functions are equivalent to likelihood ratios. This makes it possible to combine several sources of information by making the Naive Bayesian assumption that the values returned by distinct Gaussian models are independent. In particular, we can combine two functions of distributions, namely KL divergence and variation in pointwise entropy, by assuming they represent different information. As discussed in the previous section, this independence assumption can be motivated on theoretical grounds. Two results obtained by combining information from these two functions are shown in Table 5.11. In the following chapter we will also combine Gaussian models, when we develop a Maximum Description Length model for evaluating sets of substitutability relationships holding between multiple connectives.

5.6 Experiment 9: Distinguishing the order of HYPONYMY

We have so far considered the problem of distinguishing the four lexical relationships through two pseudodisambiguation tasks. One of these tasks distinguished SYNONYMY and HYPONYMY (the relationships that correlate with both higher distributional similarity and higher similarity ratings from subjects) from the other two relationships. The other task distinguished between SYNONYMY and HYPONYMY. However the HYPONYMY relationship is asymmetric, so completely determining substitutability also necessitates learning the order of the arguments to this relation. In this section we attempt this task using a number of statistical measures that have previously been proposed for learning the order of hyponymy between nouns. We also apply the new variation of pointwise entropy function, $V(p,q)$, to the task.

5.6.1 Background

Many previous studies have looked at automatically detecting hyponymy relationships between nouns. Like all other previous work discussed in this chapter, the studies are concerned with relationships between lexical types, rather than tokens. A number of these are based on the insight that there are fixed lexico-syntactic patterns that indicate hyponymy (Hearst, 1992, 1998), such as:

(5.10) "All common-law countries, **including** Canada **and** England..."
$\Rightarrow$ hyponym("Canada","common-law country") $\wedge$
hyponym("England","common-law country")

(5.11) "Bruises, wounds, broken bones, **or other** injuries ..."
$\Rightarrow$ hyponym("bruise","injury") $\wedge$ hyponym("wound","injury") $\wedge$
hyponym("broken bone","injury")

Variants of this approach have applied the same idea to new languages (Rydin, 2002), and combined a pattern matching stage with statistical information (Alfonseca and Manandhar, 2002; Caraballo, 1999; Cederberg and Widdows, 2003; Snow et al., 2004). However these pattern-based techniques are not appropriate for identifying relationships between discourse connectives. The lexico-syntactic patterns that Hearst identifies all rely on the explicit signalling of *exemplification*, and discourse connectives cannot be used as examples of other connectives. This is illustrated by the data in (5.12–5.13).

(5.12) # John left *after, including as soon as,* he finished his beer.

(5.13) # John left *as soon as or other after* he finished his beer.

There have also been attempts to predict noun hyponymy from purely statistical information. Caraballo and Charniak (1999) propose three statistics for determining which of two nouns is the more specific. Their first hypothesis is that very specific nouns are rarely modified, whereas general nouns are more commonly modified. Their second hypothesis is that the modifiers of general nouns have greater entropy, i.e. are less predictable. Their third hypothesis is simply that more general nouns occur with a greater frequency than more specific nouns. The reasoning behind this hypothesis is not clear. However it presumably relates closely to Grice's (1975) Maxim of Quantity, which states that utterances should be as informative as necessary, but not more informative than is necessary. If the writer thinks that the reader can infer the more specific aspects of meaning from the discourse context, then a more general term can safely be used. Oberlander and Knott (1995) discuss related issues concerning discourse markers, and use the term "laconic" to describe a discourse in which some inferable discourse relations remain implicit. Caraballo and Charniak test their hypotheses on hyponym–hypernym pairs from three semantic fields: food, vehicles, and occupations. Their task is to take a hyponym–hypernym pair and decide which is which. Given the symmetry of the task, the baseline performance can be considered to be 50%. They find that the predictions made by entropy and frequency give the best results, achieving about 85% and 86% accuracy, respectively (averaged across all three semantic fields).

Weeds (2002) has shown that distributional similarity measures can be used for distinguishing the order of noun hyponymy. She demonstrates that the asymmetry of Kullback-Leibler divergence can be exploited for predicting the order of hyponymy. As discussed in Section 5.4, KL divergence relates to the average surprise in replacing one lexical item with another. It is therefore expected to be greater when a hyponym is being substituted for a hypernym (so the hyponym is appropriate in only a subset of the contexts) than when a hypernym is being substituted for a hyponym. That is, we expect $D(p_{hyper}||p_{hypo}) > D(p_{hyper}||p_{hypo})$. In an experiment with 157 hyponym–hypernym pairs, Weeds finds this expectation is verified 90% of the time. Weeds et al. (2004) demonstrate a three-way correspondence between the order of hyponymy, the relative frequency of the related words, and a concept of distributional similarity based on the Co-occurrence Retrieval Model (Weeds and Weir, 2003). This model interprets distributional similarity within a framework of predicting/retrieving the co-occurrences of one distribution, given another distribution. This interpretation in terms of retrieval allows precision and recall to be calculated. High precision and/or recall is indicative of distributional

similarity, so in effect two similarity measures are defined:

- **Precision:**
$$\mathcal{P}(w_2, w_1) = \frac{\Sigma_{c \in F(w_1) \bigcap F(w_2)} I(c, w_2)}{\Sigma_{c \in F(w_2)} I(c, w_2)}$$
where $I(c,w) = \log \frac{P(c|w)}{P(c)}$ and $F(w) = \{c : I(c,w) > 0\}$.

- **Recall:**
$$\mathcal{R}(w_2, w_1) = \frac{\Sigma_{c \in F(w_1) \bigcap F(w_2)} I(c, w_1)}{\Sigma_{c \in F(w_1)} I(c, w_1)} \qquad (= sim_{\mathcal{P}}(w_1, w_2))$$
$I(c,w)$ and $F(w)$ as above.

Given two nouns n_1 and n_2, Weeds et al. predict that if $\mathcal{P}(n_2, n_1)$ is greater than $\mathcal{P}(n_1, n_2)$ (or, equivalently, $\mathcal{R}(n_2, n_1)$), then n_2 is more likely to be a hyponym of n_1, and vice versa. An empirical study using all 20,415 hyponym–hypernym pairs in WordNet 1.6 shows this prediction to be supported 71% of the time. However Caraballo and Charniak's simpler frequency-based prediction achieves comparable accuracy on the same task.

All these previous studies have concerned hyponymy between nouns. In this experiment we explore the use of previously proposed statistics to determine the order of HYPONYMY between discourse connectives. We also investigate the utility of the new variation in pointwise entropy function V for this task.

5.6.2 Hypotheses

On the basis of the above discussion, we make the following predictions regarding the probability distributions p_{hyper} and p_{hypo} of two discourse connectives standing in a HYPONYMY relationship.

Hypothesis 5.9 *The entropy of the hypernym's distribution is greater than that of the hyponym, i.e.* $H(p_{hyper}) > H(p_{hypo})$.

Hypothesis 5.10 *The KL divergence between the two distributions is sensitive to the order of the arguments, and in particular* $D(p_{hyper}||p_{hypo}) > D(p_{hypo}||p_{hyper})$.

Hypothesis 5.11 *Weeds and Weir's (2003) additive mutual information co-occurrence retrieval model is sensitive to the order in which the two distributions are supplied as arguments, and in particular* $\mathcal{P}(p_{hypo}, p_{hyper}) > \mathcal{P}(p_{hyper}, p_{hypo})$

Hypothesis 5.12 *The variance of pointwise entropy function introduced earlier is also sensitive to the order of the arguments, and in particular* $V(p_{hyper}, p_{hypo}) > V(p_{hypo}, p_{hyper})$.

We also follow Caraballo and Charniak (1999) in predicting that hyponymy affects the frequency of lexical items.

Hypothesis 5.13 *Hypernyms have greater frequencies than their hyponyms.*

5.6.3 Methodology

All 52 pairs of connectives in the HYPONYM relationship were extracted from the taxonomy and used for evaluation. We use a range of co-occurrence features that we have shown in both this chapter and the previous one to be useful for machine learning tasks. We report results using a) co-occurrences with all words in the related clauses, b) co-occurrences just with verbs, c) co-occurrences just with adverbs, and d) co-occurrences with discourse markers. (Because there are fewer parameters in this experiment than in the previous one, we can explore further options for the choice of co-occurrence type.) The distributions of these features were used to calculate the entropy of each distribution, and were also supplied as arguments to the functions for comparing distributions.

As discussed in Section 5.3, three obstacles prevent our straightforwardly estimating the relative frequencies of discourse connectives. Firstly, our methodology of obtaining example sentences from the web incorporates a web search stage that biases the sampling. Secondly, our method of identifying connectives relies on sentences being correctly parsed (or, at a minimum, relies on the parser correctly doing clause segmentation), and we cannot be sure this is equally likely for all connectives. Thirdly, in Chapter 3 we saw that the accuracy in identifying connectives varies with connectives. Therefore, we used an alternative, cruder, method of estimating connective frequencies. We only used data from the BNC, in order to avoid the problem of biased sampling from the web, and we just counted the frequencies of strings matching the surface form of each connective. That is, we did not attempt to eliminate other uses of the same phrases, such as when *and* conjoins NPs, as discussed in Chapter 3. It has been shown that the web can provide more accurate lexical frequency statistics than the BNC (Keller and Lapata, 2003). We therefore also used the web to estimate the relative frequency of discourse connectives.

Statistic	Features			
	all words	**verbs**	**adverbs**	**DMs**
$H(p_{hyper}) - H(p_{hypo})$	0.27 (0.23)	0.41 (0.20)	0.83 (1.94)	0.53 (0.43)
$D(p_{hyper}\|\|p_{hypo}) - D(p_{hyper}\|\|p_{hypo})$	0.05 (0.01)	0.08 (0.01)	-0.01 (0.03)	0.07 (0.01)
$\mathcal{P}(hypo, hyper) - \mathcal{P}(hyper, hypo)$	0.19 (0.02)	0.24 (0.04)	0.26 (0.05)	0.22 (0.03)
$V(p_{hyper}, p_{hypo}) - V(p_{hyper}, p_{hypo})$	1.80 (3.89)	1.93 (3.48)	1.57 (3.13)	2.17 (3.36)

Table 5.12: Average values (and variance) of statistics for HYPERNYMS and HYPONYMS

5.6.4 Results and discussion

Descriptive statistics for entropy, Kullback-Leibler divergence, Weeds' precision based function and variation in pointwise entropy are shown in Table 5.12. Comparison of the mean differences shows that in general the hypotheses are supported. The only exception is KL divergence using adverbs as features, for which $D(p_{hyper}||p_{hypo})$ and $D(p_{hypo}||p_{hyper})$ are about equal.

Table 5.13 summarises the performance of classifiers based on these four statistics. Weeds' precision based function, $\mathcal{P}$, gives slightly better results on average across the four features conditions. Analysis of variance shows that the choice of statistic has a significant effect on performance ($F(3,15) = 4.261, p < 0.05$), however post-hoc Tukey tests do not find any significant differences between pairs of functions. The classifier that used string frequencies in the BNC had an accuracy of 75%, however the classifier based on page hits from the web achieved 86.5% accuracy, which is about as good as the best performing classifier using distributional information (for comparison, Caraballo and Charniak (1999) report that corpus frequency gives 83.1% accuracy on predicting the direction of noun hyponymy). However these frequency-based classifiers benefited from certain substring relations between connectives. For example, *if* is a HYPERNYM of *if ever*, but it is logically necessary that the string "if" has a frequency greater than that of the string "if ever".

These results demonstrate that the statistics that have been proposed for distinguishing the order of noun hyponymy are also useful in the case of discourse connectives. From this the following three conclusions can be drawn. Firstly, $H(p_{hyper}) > H(p_{hypo})$ means that more specific connectives occur in more predictable contexts. This is because entropy is essentially a measure of randomness, so the greater the entropy of a random variable the less predictable it

	Features			
Function	**all words**	**verbs**	**adverbs**	**DMs**
H	0.654	0.827	0.731	0.865
D	0.635	0.827	0.635	0.731
$\mathcal{P}$	0.904	0.827	0.865	0.885
V	0.846	0.827	0.808	0.827

Table 5.13: Accuracy in determining the order of hyponymy

is. The entropy was lower for HYPONYMS for each of the four types of features tried, showing that HYPONYMS have more predictable co-occurrence patterns with each of verbs, adverbs and discourse markers, as well as with all words in general. This extends the findings of Caraballo and Charniak (1999) to discourse connectives. Secondly, the utility of the Kullback-Leibler divergence function extends the results of Weeds (2002) to discourse connectives. That is, the distribution of a more specific connective is a poorer approximation of the distribution of a more general connective, than the more general connective's distribution is of the more specific connective's. For example, given that *if ever* is a HYPONYM of *if*, we expect $D(if||if\ ever) > D(if\ ever||if)$. Thirdly, in Weeds' co-occurrence retrieval based framework, a more specific connective has greater precision in retrieving a more general connective's occurrences than it has recall. The results also demonstrate that the new variation of pointwise entropy function performs about as well as the previously proposed statistics on this task.

Given that both entropy and Kullback-Leibler divergence are useful predictors of the order of both noun hyponymy and discourse connective HYPONYMY, it is possible that variation in pointwise entropy might also be a useful predictor for the order of noun hyponymy. However investigating this hypothesis lies beyond the scope of this thesis.

5.7 Summary

The concepts of lexical similarity and substitutability are of central importance to psychology, artificial intelligence and computational linguistics. In this chapter we demonstrated a three way correspondence between data sources of quite distinct types: distributional similarity scores obtained from lexical co-occurrence data, substitutability judgements made by trained analysts, and the similarity ratings of naive subjects. The convergence of different

sources of data provides evidence that the phenomena are robust. This suggests the possibility that the similarity and substitutability of connectives might have some cognitive reality in the structuring of the mental lexicon, in the same way that WordNet claims to be psychologically plausible.

We presented experiments indicating that the degree of similarity of pairs of discourse connectives can be quantified, extending previous findings for nouns or verbs. In particular, two sources of evidence were found to support this: (1) human subjects show significant agreement when rating the similarity of discourse connectives, and (2) subject ratings of similarity correlate with the predictions of a distributional model. For the latter, we only found a significant correlation when we used co-occurrences of the discourse connectives with other discourse connectives or with discourse adverbials. The interpretation of discourse adverbials co-occurring with discourse connectives is known to be complex (Webber et al., 2003), but this result suggests that similar connectives have similar patterning of co-occurrences with adverbials.

In this chapter we also introduced a new variance-based function of two distributions and demonstrated its utility in automatic lexical acquisition. Many previous functions have measured distributional (dis)similarity, and combining such functions with a variance-based one allows a two-dimensional view of the data to be obtained. As a result, it can be useful to combine the predictions of the new function with those of previous functions, either by making Bayesian independence assumptions or by constructing ensembles.

In the following chapter we continue to explore the acquisition of substitutability relationships. However we progress from considering pairs of connectives to considering sets of many connectives, and attempt to learn the entire set of pairwise relationships that hold between them.

Chapter 6

Developing taxonomies of discourse connectives

The previous chapter concerned the machine learning of relationships that hold between pairs of discourse connectives. In particular, it considered methods for predicting the similarity of connectives, and for learning substitutability relationships between connectives. However learning substitutability was hampered by the high prior likelihood of connectives being EXCLUSIVE. As a result, we only attempted a pseudodisambiguation task that made prior likelihoods irrelevant. Although this task was useful for exploring techniques for predicting substitutability, the task itself did not actually constitute predicting substitutability relationships.

In this chapter we show that the effects of prior likelihoods can be overcome by modelling the global structure of the lexicon of discourse connectives. The demonstration of this is one of the four main primary contributions of the thesis, as outlined in Chapter 1. The model that we develop will require that the set of pairwise relationships between discourse connectives is globally consistent. For example, given three discourse connectives A, B and C, Figure 6.1 gives an example of a set of relationships that is globally consistent, as well as a set that is inconsistent. When we come to automatically predicting relationships between connectives, we will utilise this requirement of consistency to constrain our search space. We will do this by automatically constructing entire taxonomies of connectives, rather than just considering pairs in isolation. This model will also penalise taxonomies in which a single relationship (e.g. EXCLUSIVE) holds between all pairs of connectives. This will help to overcome the problems encountered in the previous chapter.

We proceed by first briefly reviewing the role of lexical taxonomies in computational lin-

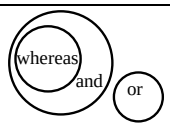
Pairwise relationships	**Consistent?**	**Venn diagram**
HYPONYM(whereas,and), EXCLUSIVE(whereas,or), EXCLUSIVE(and,or)	Yes	
HYPONYM(whereas,and), CONT. SUBS.(whereas,or), EXCLUSIVE(and,or)	No	*Not possible*

Figure 6.1: Consistent and inconsistent sets of substitutability relationships

guistics, and considering previous attempts to induce them automatically. We then introduce a new model of taxonomies of discourse connectives based on the Minimum Description Length principle (cf. Rissanen, 1978). This model is then applied in an experiment into automatically extending an existing taxonomy with additional connectives.

6.1 Background

In this section, we will use the term *taxonomy* broadly, including a range of data structures that also go by names such as *ontologies* and *inheritance hierarchies*. One thing these have in common is that they constrain the relationships that may hold between sets of items. As the most extreme example of this, if two items occupy the same position in a taxonomy, then they must both be related to all other items in the taxonomy in the same way. Subordination relationships also produce strong constraints on which pairwise relationships are possible. For example, if X is subordinate to Y then there cannot be a third item that is subordinate to X but not subordinate to Y.

Taxonomies have been widely used in computational linguistics because of the practical benefits they convey. Subordination relationships allow generalisations to be captured, and in the process redundancy in the lexicon can be reduced or eliminated (Daelemans et al., 1992; Briscoe et al., 1993). By enabling a compact representation, the memory requirements for storing the lexicon are reduced. Benefits are also obtained when using structured lexicons for processing tasks. Tree-like structures (and, more generally, directed acyclic graphs) can enable efficient searching of the lexicon, and the explicit representation of generalisations can be utilised for smoothing statistical language models. Because of these benefits, taxonomies

have been widely used in computational linguistics. WordNet is a commonly used taxonomy representing lexical semantic relations between nouns, verbs, adjectives and adverbs (Miller, 1990). Lexicalised theories of grammar have used inheritance hierarchies to capture grammatical generalisations. This application has been explored to the greatest degree within the framework of Head-Driven Phrase Structure Grammar (Pollard and Sag, 1987, 1994), however it has also been applied to Lexicalised Tree Adjoining Grammar (Vijay-Shanker and Schabes, 1992) and Categorial Grammar (Baldridge, 2002; Villavicencio, 2001). Applications such as these have led to the development of a number of taxonomy development environments, which provide support for making changes to a taxonomy, and can ensure the internal consistency of a taxonomy.

However the manual construction of large taxonomies is still a difficult task. In particular, they are difficult to maintain, update and check for consistency. This is partly due to the fact that even minor changes can have far reaching effects. For example, changing which relationships are possible between two lexical items can affect the relationships not only between these items but also between other items in the taxonomy. It is easy to overlook some of the consequences of making such a change, and this problem increases as the size of the taxonomy grows. For these reasons, automated and semi-automated (i.e. computer-assisted) development of taxonomies are of interest (e.g. Sporleder, 2004a). Possible applications include:

1. Automatically extending a manually constructed taxonomy;

2. Constructing a taxonomy from scratch; and

3. Assisting a human to extend or construct a taxonomy.

The first and second of these tasks can both be attempted using either data-driven or data-free approaches, depending on how much linguistic information is already known. If the linguistic features of each lexical item are known beforehand then corpus data may not be required. For example, Petersen (2001) uses a set-theoretical approach borrowed from Formal Concept Analysis (Ganter and Wille, 1999) to create inheritance hierarchies for words whose linguistic features are known. Cimiano et al. (2004) adopt a similar approach to Petersen, but induce linguistic features from a corpus rather than taking them for granted. Both these approaches produce hierarchies that are free of redundancy, however Sporleder (2004b) has argued that such hierarchies are not always the most linguistically plausible. Instead, Sporleder (2004a; 2004b) combines machine learning with set-theoretic methods to induce inheritance hierarchies with similar "shape" to manually constructed hierarchies. Villavicencio (2001) learns

hierarchies of grammatical categories in the framework of categorial grammar, from input consisting of pairings of sentences and semantic representations. Her system searches for the hierarchy which is the most compact, in the sense that the maximal amount of attribute–value structures at the leaf nodes are inherited from parent nodes. Work on automatically extending taxonomies has been focused on extending the WordNet ontology. Hearst and Schütze (1996) assign 27 new words into disjoint categories derived from WordNet, using a model based on lexical co-occurrence statistics. Alfonseca and Manandhar (2002) extend WordNet with terms from the novel *The Lord of the Rings* (Tolkien, 1954), e.g. *hobbit*, using both distributional similarity and templates of lexical patterns (Hearst, 1998). Widdows' (2003) approach to extending WordNet combines distributional similarity with part-of-speech information.[1]

Finally, there are semi-automated construction tasks, under which we include cases where the system produces a taxonomy that is intended from the outset to be post-edited by a human. (Post-editing is more common in the field of Machine Translation (Knight and Chander, 1994; Allen, 2003), however it can also be applied to taxonomy construction.) Such semi-automatic tasks have much in common with fully automatic taxonomy construction. However the quality of assistance can be improved by the system's offering the user additional information about its predictions. For example, the user may want to know the reason behind a certain decision (in which case decision trees can be informative classifiers), or the system might return confidence scores accompanying each judgement.

In the first experiment reported below, we attempt the first of the three tasks considered above: extending an existing taxonomy. In particular, we will add new connectives to a taxonomy representing substitutability relationships. We choose the first task because success on this task is a prerequisite for success on the second. That is, if we cannot extend an existing taxonomy from scratch then there is little hope that we can create one from scratch. In addition, when extending a taxonomy we have an existing (although incomplete) taxonomy from which we are able to estimate the prior likelihoods of the different substitutability relationships. Before presenting the experiment, we first introduce a statistical model of taxonomies that takes into account both the prior likelihood of the taxonomy and how well it explains the data. This model provides a mechanism for choosing between different extensions of the original taxonomy.

[1] Although it is does not result in true taxonomies, there has also been work on the related task of clustering lexical items based on corpus evidence (for example Pereira et al., 1993; Brew and Schulte im Walde, 2002; Li, 2002; Cimiano et al., 2004).

6.2 Modelling taxonomies

Our modelling of taxonomies is within the Minimum Description Length (MDL) framework (Rissanen, 1978). MDL is a principle of empirical model evaluation based on information theory. It states that the best model for some data is the one which requires the minimum number of bits in order to encode both the model itself and the data as observed through the model (Quinlan and Rivest, 1989; Li and Abe, 1998). In our case, the taxonomies we wish to model represent substitutability relationships between connectives. Interpreted in the MDL framework, taxonomies are evaluated according to both a) their prior likelihood, and b) how well the taxonomy explains the data. For a taxonomy $\mathbb{T}$ and data *data*, the total description length L is:

$$L(\mathbb{T}, data) = L(\mathbb{T}) + L(data|\mathbb{T}) \tag{6.1}$$

To calculate this value, we will exploit the fact that MDL has a Bayesian interpretation which relates description lengths L to probabilities P via the equation $L = -\log_2 P$. This follows from the assumption that more probable taxonomies (or, more generally, models) can be encoded using fewer bits. We therefore proceed by deriving probability models for $P(\mathbb{T})$ and $P(data|\mathbb{T})$. We will refer to these probabilities as the *prior* and the *posterior*, respectively. In deriving these probability models, we will consider a taxonomy to be equivalent to the set of pairwise relationships that it contains. That is, it is just the logical content of the taxonomy, rather than any organisational aspects, that concerns us. Our approach lets us express our model in terms of pairwise relationships, which we treat as elementary units. This will enable us to develop an efficient implementation of the model. Our approach is in sharp contrast to studies which are concerned with the appropriate organisation of a predetermined set of relationships between lexical items (for example as defined by linguistic features, e.g. Petersen, 2001; Sporleder, 2004a).

We proceed by deriving a formula for the prior probability (i.e. $P(\mathbb{T})$) that takes the global distribution of the different substitutability relationships into account. It will, for example, penalise taxonomies which posit a single type of relationship (e.g. EXCLUSIVE) holding between all pairs of connectives in the taxonomy, on the grounds that they contain relationships in unexpected ratios. We then derive a method for estimating the posterior (i.e. $P(data|\mathbb{T})$) that compares the substitutability relationships posited between pairs of connectives with the distributional similarity of those pairs.

6.2.1 Calculating the prior

To calculate the prior, we will consider a taxonomy $\mathbb{T}$ to be equivalent to the set of its pairwise relationships between connectives. In doing so, certain aspects of the taxonomy will be ignored. For example, we do not directly take into account the depth of subordination in the taxonomy. However, it has been observed, for example, that lexical inheritance hierarchies seldom go more than ten levels deep (Miller, 1990). In addition, although Knott (p. 80) observes that his taxonomy often has hyponymic chains of length two or three, the maximum chain length in his taxonomy is only four (e.g. *so–therefore–as a result–thereby*). So one might, in principle, want to model the length of such chains of subordinates as a soft constraint. We also do not take into account the total number of separate hyponyms that connectives have. However Knott (p. 80) observes that in his taxonomy *and* has thirty hyponyms, and it is almost inconceivable that any connective that Knott has not considered would have more.

In order to express the model succinctly, it will be useful to introduce some notation. We will use $\langle rel, X, Y \rangle$ to denote that connectives X and Y are in relationship *rel*, where *rel* is one of SYNONYM, EXCLUSIVE, etc. And we will use $\langle rel, X, Y \rangle \in \mathbb{T}$ to indicate that taxonomy $\mathbb{T}$ implies that the relationship *rel* holds between X and Y. To qualify as a taxonomy, we require that $\mathbb{T}$ be complete, consistent and free of redundancy. That is, for all connectives X, Y in the taxonomy, there must be a unique *rel* such that either $\langle rel, X, Y \rangle \in \mathbb{T}$ or $\langle rel, Y, X \rangle \in \mathbb{T}$. Examples of sets of relationships that violate these three conditions are given below.

- $\{\langle \text{EXCLUSIVE}, X, Y \rangle, \langle \text{EXCLUSIVE}, X, Z \rangle\}$ is incomplete, since there is no relationship between Y and Z.
- $\{\langle \text{EXCLUSIVE}, X, Y \rangle, \langle \text{SYNONYM}, X, Y \rangle\}$ is inconsistent, since two different relationships hold between X and Y.
- $\{\langle \text{EXCLUSIVE}, X, Y \rangle, \langle \text{EXCLUSIVE}, Y, X \rangle\}$ contains redundancy, since the same relationship is stated twice.

Note that by "redundancy", we here simply mean the formal redundancy whereby pairwise relationships are repeated. This is different to the structural redundancy discussed by Sporleder (2004b).

To calculate $P(\mathbb{T})$ (and also, later, $P(data|\mathbb{T})$) we will consider products over all pairwise relationships $\langle rel, X, Y \rangle$ in the taxonomy $\mathbb{T}$. To calculate $P(\mathbb{T})$, we use the following multino-

# connectives	1	2	3	4	5	6
# sets of relationships	1	5	125	15,625	9,765,625	30,517,578,125
# consistent sets	1	5	54	968	29,709	*memory error*

Table 6.1: Sets of relationships between N connectives

mial model:

$$P(\mathbb{T}) \propto M \prod_{\langle rel,X,Y\rangle \in \mathbb{T}} P(rel) \tag{6.2}$$

where (i) $P(rel)$ is the prior probability of any two connectives being in the relationship *rel* (which will be estimated empirically), and (ii) M is a multinomial coefficient that ensures that the most likely taxonomy contains numbers of each pairwise substitutability relationship in proportion to their prior probabilities (for a comparison of multinomial models to naive Bayes models, see Eyheramendy et al. (2003)). If the numbers of each type of substitutability relationship in $\mathbb{T}$ are given by N_{syn}, N_{hyp}, N_{ex} and N_{cont}, then:[2]

$$M = \frac{(N_{syn} + N_{hyp} + N_{ex} + N_{cont})!}{N_{syn}!N_{hyp}!N_{ex}!N_{cont}!} \tag{6.3}$$

The multinomial model (6.2) is defined over all sets of pairwise substitutability relationships. However we are only interested in calculating the probabilities of sets of relationships that are logically consistent. To do this, we assign zero probability to sets of relationships which contradict the logic of set theory. For example the set $\{\text{SYNONYM}(A,B), \text{EXCLUSIVE}(A,C), \text{SYNONYM}(B,C)\}$ is inconsistent and so is assigned zero probability. By doing this we reduce the total probability mass, and so in principle we would want to correct for this. In order to do so, we would have to calculate the number of consistent sets of pairwise relationships, as well the total number of (possibly inconsistent) sets. For N connectives, the latter value is simply $5^{\binom{N}{2}}$, as there are $\binom{N}{2}$ pairs of connectives, and five possible relationships (treating the different orderings of the HYPONYMY relation as distinct). However calculating the former value is more challenging. A programme was written to calculate this value by generating all sets and checking their consistency. The values obtained for small N are shown in Table 6.1, however for N equal to 6 or above the memory requirements were too great. As a result, it is not in general feasible to calculate how much probability mass is unaccounted for by our model, and

[2]This formula easily generalises to cover taxonomies representing different types of relationships. The general formula is: $M = \frac{(\sum_{rel} N_{rel})!}{\prod_{rel} N_{rel}!}$.

so correcting for this missing mass is also not feasible.[3] Fortunately, the missing mass does not affect the likelihood ratios of taxonomies with the same number of connectives, so for the purposes of the experiments below the missing probability mass can be ignored.

6.2.2 Estimating the posterior

The data that our model of taxonomies aims to explain are the co-occurrence distributions of each connective. However, unlike some related work applying MDL to lexical co-occurrences (Li and Abe, 1998), we apply our model to distributions of co-occurrences, so that our taxonomy aims to explain *relationships between co-occurrence distributions*. For the time being we will remain agnostic as to what types of relationships these might be, but one obvious choice is the distributional similarity of connectives. There are two reasons for this divergence from Li and Abe's approach. Firstly, our task differs substantially from theirs in that they learn generalisations from a given taxonomy, whereas our immediate goal here is to estimate the likelihood of a taxonomy using empirical data. Secondly, by modelling relationships between connectives we are able to exploit the correspondences between substitutability and distributional similarity that we illustrated in the previous chapter. For each pair X,Y in the taxonomy, the data will include the ordered pair of X and Y's substitutability relationship and their distributional relationship:

$$data \equiv \{(rel, f(X,Y)) : \langle rel,X,Y\rangle \in \mathbb{T}\} \tag{6.4}$$

where f is some function of the distributional representations of connectives X and Y. To make the following exposition more concrete, we shall assume for the time being that f is the Kullback-Leibler divergence function, so that the empirical data the taxonomy aims to explain are the divergences between the distributions of connectives.

To estimate the probability of the data, we assume that the likelihood of observing a given distributional divergence $D(X||Y)$ between X and Y is dependent only on the substitutability of X and Y. That is:

$$P(data|\mathbb{T}) = P(\{(rel,D(X||Y)) : \langle rel,X,Y\rangle \in \mathbb{T}\}|\mathbb{T}) \tag{6.5}$$

$$\approx \prod_{\langle rel,X,Y\rangle \in \mathbb{T}} P(D(X||Y)|\langle rel,X,Y\rangle) \tag{6.6}$$

[3] It may be that the number of consistent sets can be expressed using a succinct mathematical formula, making empirical calculation redundant. However a search of an on-line encyclopedia of integer sequences (`http://www.research.att.com/~njas/sequences/`) did not return any sequences beginning with the first few values of the sequence shown in Table 6.1.

Note that this assumption is a simplification, since the distributional divergences are not in fact independent. This can be seen clearly by supposing that X and Y have a distributional divergence of 0, and that Y and Z have a divergence of 1. Then, since X and Y have the same distributions, X and Z must also have an empirical divergence of 1, irrespective of their substitutability. In the general case, the definition of Kullback-Leibler divergence places constraints on the divergences that are possible between three or more connectives. The degree to which this simplification affects the predictions of the model will be explored in the experiment described below.

To estimate each of the multiplicands in (6.6), we use a Gaussian model of the distributional divergences corresponding to each substitutability relationship. That is, for each relationship *rel*, we take all pairs in that relationship and calculate the mean μ_{rel} and standard deviation σ_{rel} of their distributional divergences. From these, a Gaussian function $n(\cdot;\mu,\sigma)$ can be calculated. (In the experiments described below, we in fact use a log-normal model of the KL divergences, due to their being right-skewed.) Gaussian models are continuous functions and so cannot be used to calculate probabilities of individual values. However, values of a finite number of Gaussians at a particular point are proportional to posterior likelihoods:

$$P(D(X||Y)|\langle rel, X, Y\rangle) \propto n(D(X||Y); \mu_{rel}, \sigma_{rel}) \tag{6.7}$$

where μ_{rel} and σ_{rel} are the mean and standard deviation of KL divergences of all pairs of connectives in relationship *rel*. As a result, the posterior likelihood can be estimated by combining (6.6) and (6.7).

We now give an illustrative example of how calculation of the prior and posterior probabilities can be used to decide which of alternative taxonomies is most likely. Consider the three sets of relationships between connectives given in Figure 6.2. The third set is ruled out immediately on the grounds that it is logically inconsistent. Of the remaining two taxonomies, the second has a greater prior likelihood than the first, due to CONTINGENTLY SUBSTITUTABLE having a greater prior probability than HYPONYMY (all numerical values in Figure 6.2 have been invented for illustrative purposes). However the co-occurrence distributions have a greater posterior likelihood when the first taxonomy is assumed (and hence $L(data|\mathbb{T})$) is lower). In fact, this better fit to the data causes the overall description length of the first taxonomy to be lower than that of the second (the final column of the Figure), and hence the first taxonomy is preferable.

As noted above, instead of $D(X||Y)$ we might in practice want to use alternative functions $f(X,Y)$ of the distributions of connectives. For example, we might wish to use the variation

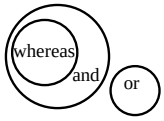
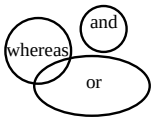
Pairwise relationships	**Consistent?**	**Venn diagram**	$L(\mathbb{T})$	$L(data\|\mathbb{T})$	$L(\mathbb{T},data)$
HYPONYM(whereas,and), EXCLUSIVE(whereas,or), EXCLUSIVE(and,or)	Yes		3.7	1.5	5.2
EXCLUSIVE(whereas,and), CONT. SUBS.(whereas,or), EXCLUSIVE(and,or)	Yes		2.1	3.8	5.9
HYPONYM(whereas,and), CONT. SUBS.(whereas,or), EXCLUSIVE(and,or)	No	*Not possible*	—	—	—

Figure 6.2: Illustrative example of how the MDL model is applied

in pointwise entropy function $V(X,Y)$ introduced in the previous chapter. If so, the procedure remains the same: the mean and standard deviation of f are calculated for each substitutability relationship, and these are used to construct Gaussian functions for estimating the fit of the taxonomy to the empirical data. In fact, we can easily generalise the method in order to use multiple functions, say both $D(X||Y)$ and $V(X,Y)$, of the distributions of X and Y, thus enabling a richer representation of the data. For example, our data could consist of triples of substitutability relationships, distributional divergences, and variations in pointwise entropy:

$$data \equiv \{(rel, D(X||Y), V(X,Y)) : \langle rel, X, Y\rangle \in \mathbb{T}\} \tag{6.8}$$

We will refer to such representations of the data as "compound". Non-compound representations of the data will be called "simple". We will then assume that D and V are independent of each other, allowing us to simply multiply the individual probabilities:

$$P(data|\mathbb{T}) = P(\{(rel, D(X||Y), V(X,Y)) : \langle rel, X, Y\rangle \in \mathbb{T}\}|\mathbb{T}) \tag{6.9}$$

$$\approx \prod_{\langle rel,X,Y\rangle \in \mathbb{T}} P(D(X||Y)|\langle rel, X, Y\rangle) \prod_{\langle rel,X,Y\rangle \in \mathbb{T}} P(V(X,Y)|\langle rel, X, Y\rangle) \tag{6.10}$$

Together, the prior and posterior probability models allow the description length of a taxon-

omy to be calculated. The following experiment applies description lengths to the task of inserting new connectives into an existing taxonomy.

6.3 Experiment 10: Extending a taxonomy of connectives

The probability model of taxonomies presented in the previous section provides a principle for deciding how to extend taxonomies of discourse connectives. One taxonomy is to be preferred to a rival if it has a shorter description length. There is a close relationship here to Ockham's Razor: given two theories of the data, the one which is simpler (i.e. has the shorter description length) is to be preferred. Our model has the desirable property that $L(\mathbb{T})$ is minimal when the frequency of each type of substitutability relationship is in proportion to its prior likelihood. However the overall description length $L(\mathbb{T}, data)$ also takes the empirical fit of the taxonomy to the data into account. In this experiment we apply the model to the task of extending an existing taxonomy automatically. The methods applied to the task might be used to extend Knott's taxonomy of connectives. Furthermore, as discussed above, being able to extend a taxonomy can be considered a prerequisite for constructing a taxonomy from scratch. Thus this experiment also has potential consequences for the creation of new taxonomies of discourse connectives, for example, of connectives in languages other than English.

6.3.1 Hypothesis

In the previous chapter we saw that the high prior likelihood of the EXCLUSIVE relationship was an obstacle to predicting other relationships with a high degree of precision. However our MDL-based model takes into account the global distribution of substitutability relationships. In particular, the multinomial term M in equation (6.2) creates a bias towards taxonomies containing a mixture of different substitutability relationships.

Hypothesis 6.1 *Modelling the global distribution of relationships in a taxonomy improves performance in automatically extending an existing taxonomy.*

6.3.2 The task

Our task in this experiment is to insert new connectives into an existing taxonomy of connectives. Since we are equating a taxonomy with the set of pairwise relationships it contains, this equates to predicting the substitutability relationships that hold between the new connective

and the connectives already in the taxonomy. However the requirement that the taxonomy be logically consistent constrains our prediction of pairwise relationships.

6.3.3 Methodology

We take a manually constructed taxonomy of connectives as a gold standard. We then remove a single connective from the taxonomy, and attempt to re-insert it in its original position. This methodology has previously been used for researching taxonomy extension (e.g. Widdows, 2003), and has the benefit of making evaluation easier since the correct result is given by the gold standard. In our case, we attempt to find the taxonomy $\mathbb{T}'$ such that (a) $\mathbb{T}'$ is consistent with the subtaxonomy formed when the connective was removed, and (b) of all consistent taxonomies, $\mathbb{T}'$ has the minimum description length. That is, we must solve the following equation:

$$\mathbb{T}' = \arg\min_{T} L(T, data) \tag{6.11}$$

with the constraint that the T can differ from the original taxonomy only in the relationships involving the removed connective. We could do this for just a selection of connectives, however to provide a more rigorous evaluation we will do it for all of them. Our procedure is thus a type of leave-one-out cross-validation.

The gold standard taxonomy used is the one introduced in Chapter 3. It contains 80 connectives. We also re-used the co-occurrence data introduced in that chapter to calculate the Kullback-Leibler divergences and the variation in pointwise entropy ($V(X,Y)$) for each pair of connectives.

The prior probabilities of each substitutability relationship were re-estimated during each stage of the task. That is, after each connective was removed from the taxonomy, the remaining 79 connectives were used to give empirical estimates of the prior probabilities of SYNONYM, HYPONYM, etc. These priors were then used when calculating $L(\mathbb{T})$ in order to re-insert the extracted connective. This procedure has similarities both with the leave-one-out methodology used in Chapter 4, and with the leave-one-out resampling used to calculate inter-subject agreement in Chapter 5.

The space of taxonomies containing 80 connectives is enormous. Even the subspace in which the relationships between 79 of these are fixed is too large to search completely. To get an idea of how quickly the space of insertion positions grows, consider the following simple subtaxonomy (from Knott, 1996), containing just six pairwise relationships between four

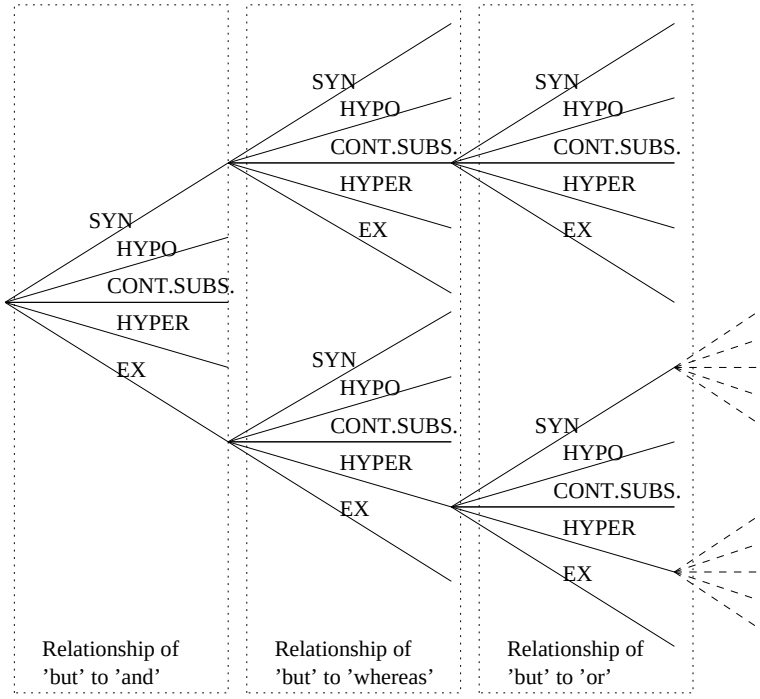


Figure 6.3: Application of beam search to taxonomy extension

connectives:[4]

(6.12) $\mathbb{T} = \{\langle$CONT. SUBS.,and,or$\rangle$, $\langle$CONT. SUBS.,and,while$_2\rangle$,
$\langle$HYPONYM,whereas,while$_2\rangle$, $\langle$HYPONYM,whereas,and$\rangle$,
$\langle$EXCLUSIVE,while$_2$,or$\rangle$,$\langle$EXCLUSIVE,whereas,or$\rangle$ $\}$

Even for a taxonomy this small there are 46 logical possibilities for how a new connective might be inserted. Since the number of possible taxonomies is exponential in the number of connectives, it is difficult to find an exact solution to the Equation 6.11. As a result we use beam search to make searching the insertion positions feasible. To do this, we decompose the insertion of a connective into a taxonomy into a series of decisions. Each decision involves determining a single pairwise relationship between the new connective being inserted and one of the connectives already in the taxonomy. For example, if we are inserting the connective *but* into the taxonomy in (6.12) then the first decision might constitute determining its substitutability with *and*. Figure 6.3 provides a visual illustration of this via a search tree. The second decision might determine *but*'s substitutability with *whereas*, the third with *or*, and

so on. After each decision, we prune our list of candidates using a fixed beam size, and the experiment was re-run with various beam sizes to determine whether this had a major effect. To determine the order in which new pairwise relationships will be decided, we first sort all connectives already in the taxonomy according to their distributional similarity to the connective being inserted. That is, the relationship of the inserted connective to its distributionally most similar connective is determined first, followed by its relationship to the connective it is distributionally second most similar to, and so on. In doing so, the consistency of the sets of relationships is always taken into account; that is, inconsistent search paths are pruned.

Technically, this application of beam search requires extending the MDL-based model of taxonomies we introduced earlier. The model was previously defined only for fully specified taxonomies, so that all pairwise relationships between connectives are determined. However the beam search requires re-evaluation after the addition of each new pairwise relationship between the connective being inserted and previous members of the taxonomy. Fortunately, the fact that both $P(\mathbb{T})$ and $P(data|\mathbb{T})$ are expressed as products over pairwise relationships makes it trivial to extend the model. All we do is extend the products in (6.2) and (6.6) to arbitrary sets $\mathcal{S}$ of pairwise substitutability relationships. That is:

$$P(\mathcal{S}) \propto M \prod_{\langle rel,X,Y\rangle \in \mathcal{S}} P(rel) \tag{6.13}$$

$$P(data|\mathcal{S}) \approx \prod_{\langle rel,X,Y\rangle \in \mathcal{S}} P(D(X||Y)|\langle rel,X,Y\rangle) \tag{6.14}$$

The model and search strategy involve a number of parameters. Different parameter settings were experimented with to explore the effects of each parameter. The parameters and their settings were:

Beam size: fixed widths of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100, 200, 500 or 1000 were tried.

Co-occurrence feature: co-occurrences with discourse markers, with verbs, or with all open-class words.

Distributional function: Kullback-Leibler divergence ($D(X||Y)$) or variation in pointwise entropy ($V(X,Y)$).

Symmetrisation: To construct the Gaussian functions we need to calculate the mean μ_{rel} of the distributional function applied to all pairs in relationship *rel*. However, the distributional

[4]Knott uses "while$_2$" to denote the contrastive sense of *while*.

functions $D(X||Y)$ and $V(X,Y)$ are asymmetric, so it is not clear in which order the arguments should be applied. To overcome this, the distributional function was calculated with arguments in both possible orders and one of the following symmetric functions applied:

- "Average", e.g. $\frac{D(X||Y)+D(Y||X)}{2}$.
- "Average of logarithms", e.g. $\frac{\log(D(X||Y))+\log(D(Y||X))}{2}$. This symmetrisation is more appropriate than the previous one for distributional functions with a right skewed distribution.
- "Difference squared", e.g. $(V(X,Y)-V(Y,X))^2$. This symmetrisation aims to exploit the discovery in Chapter 5 that the asymmetry of $V(p,q)$ can be used to predict substitutability.
- "Difference squared over average", e.g. $(V(X,Y)-V(Y,X))^2/\frac{V(X,Y)+V(Y,X)}{2}$. Similar to the previous one, except normalising for magnitude.

Type of data representation: As discussed in Section 6.2.2, the representation *data* of pairs of distributions can be be either simple or compound, where by "compound" we mean that it combines more than one function of the distributional representations. We experimented with three types of compound representations:

- Compound$_{cooc}$: those which combine different types of lexical co-occurrences (e.g. co-occurrences with verbs and co-occurrences with discourse markers), but use the same distributional function.
- Compound$_{func}$: those which combine different distributional functions of a single type of lexical co-occurrences.
- Compound$_{both}$: those which combine both different functions and different types of co-occurrences.

6.3.4 Implementation issues

The expression of the model in terms of products over pairwise relationships allows for an efficient implementation of the search strategy. In particular, dynamic programming techniques allow previously calculated probabilities to be re-used. At each stage of the beam search, an additional pairwise relationship $\langle rel_{i+1}, X_{i+1}, Y_{i+1}\rangle$ is added to the set $\mathcal{S}_i$ of pairwise relationships already posited, resulting in a new set $\mathcal{S}_{i+1} = \mathcal{S}_i \bigcup \{\langle rel_{i+1}, X_{i+1}, Y_{i+1}\rangle\}$. The prior and

posterior probabilities of $\mathcal{S}_{i+1}$ can be expressed concisely in terms of those of $\mathcal{S}_i$:

$$P(data|\mathcal{S}_{i+1}) \approx \prod_{\langle rel,X,Y\rangle \in \mathcal{S}_{i+1}} P(f(X,Y)|\langle rel,X,Y\rangle) \tag{6.15}$$

$$= P(data|\mathcal{S}_i) \times P(f(X_{i+1},Y_{i+1})|\langle rel_{i+1},X_{i+1},Y_{i+1}\rangle) \tag{6.16}$$

$$P(\mathcal{S}_{i+1}) \propto M_{i+1} \prod_{\langle rel,X,Y\rangle \in \mathcal{S}_{i+1}} P(rel) \tag{6.17}$$

$$= P(\mathcal{S}_i) \times M' P(rel_{i+1}) \tag{6.18}$$

where $M' = \frac{M_{i+1}}{M_i}$ corrects the multinomial term, and can be efficiently calculated as follows. Let N_{rel} be the frequency of relationship *rel* in $\mathcal{S}$, then:

$$M' = \frac{M_{i+1}}{M_i} \tag{6.19}$$

$$= \frac{\frac{(1+\sum_{rel} N_{rel})!}{(1+N_{rel_{i+1}})! \prod_{rel \neq rel_{i+1}} N_{rel}!}}{\frac{(\sum_{rel} N_{rel})!}{\prod_{rel} N_{rel}!}} \tag{6.20}$$

$$= \frac{1+\sum_{rel} N_{rel}}{1+N_{rel_{i+1}}} \tag{6.21}$$

In practice, we perform all calculations in terms of log-probabilities (≡ 'lengths'), in order to minimise rounding errors.

6.3.5 Evaluation metrics

We will use several metrics to evaluate performance on this task. Since we are equating taxonomies with the sets of pairwise relationships they represent, our first metric is simply the accuracy in predicting these relationships. In this experiment we are attempting to insert a connective into the correct position in the taxonomy, i.e. insert a connective so that it stands in the correct relationships to other connectives. So if we insert the new connective such that it has correct relationships to half of the connectives already in the taxonomy, we would achieve an accuracy of 50%. Since we iteratively remove and re-insert all 80 connectives in the gold standard taxonomy, we evaluate $80 \times 79 = 6320$ predicted relationships.

While accuracy is a commonly-used metric that is easy to understand, it is not the most informative measure for lexical acquisition tasks. The accuracy metric only considers whether an item is classified correctly; it is blind as to the size of the class the item belongs to, and to what types of errors are being made. This may not be important for classification tasks where the classes are of roughly equal size, however that is not the case here. Almost 70% of all pairs of connectives are EXCLUSIVE, while less than 1% of pairs are SYNONYMS. If a classifier

correctly predicted a pair of connectives to be EXCLUSIVE then it has not done much work. All it has done is predict what might be considered the default relationship. Conversely, if another classifier correctly predicts a pair to be SYNONYMS then it has performed a much harder task. This second classifier can be considered to have acquired more knowledge than the first one, and this can be formalised using ideas from information theory.

Kononenko and Bratko (1991) propose an information-based criterion for evaluating classifiers that has been applied to lexical acquisition tasks such as classifying proper names (Cucchiarelli et al., 1998), learning lexical categories (Durieux et al., 1999), and integrating semantic lexicons with domain ontologies (Basili et al., 2004). Kononenko and Bratko define a *Relative Information score* measure, I_r, which has the following key properties:

1. The correct classification into a more probable class is rewarded less than the correct classification into a less probable class.

2. The incorrect classification of an item belonging to a more probable class is penalised more than the incorrect classification of an item from a less probable class.

3. If all items are classified correctly, then $I_r = 1$.

4. If a classifier incorrectly classifies all items, then $I_r < 0$, although its precise value will depend on the class prior probabilities. (The greater the entropy of the prior probabilities, the less the classifier is penalised for failure.)

The Relative Information score is defined in terms of the amount of *Obtained Information* (I_o) and the amount of *Misleading Information* (I_m). Suppose an item belongs to class C. If it is correctly classified then it contributes $-\log_2 P(C)$ to the sum of Obtained Information. If however it is incorrectly classified then it contributes $-\log_2(1-P(C))$ to the sum of Misleading Information. The *Average Information* score I_a is obtained by deducting the misleading information from the obtained information and dividing by the total number of items.

$$I_a = \frac{I_o - I_m}{N} \tag{6.22}$$

The Relative Information score I_r normalises the Average Information by taking into account the entropy E of the class prior probabilities.

$$I_r = I_a / E \tag{6.23}$$

This normalisation ensures that the maximum possible value of I_r is 1. In automatic lexical acquisition, one might arguably care most about optimising either the amount of Obtained Information or about the Relative Information score, and so we will report both. Obtained Information measures how much lexical knowledge (measured in bits) has been learnt, without worrying about what kinds of mistakes were made in the process. It is somewhat like a recall score in this respect. Conversely, the amount of Misleading Information relates to precision. In contrast, Relative Information is more like an F-score in that it takes both into account.

Finally, the kappa statistic (κ) will also be used as an evaluation metric. Although this statistic is more commonly used within NLP for assessing inter-annotator agreement (Carletta, 1996), it can also be used to compare the performance of a classifier with a gold standard classification (Teufel and Moens, 2002). The kappa statistic is useful as an evaluation measure because it takes into account the degree of agreement with the gold standard that can be expected purely by chance. If we let this value be $P(E)$, and the accuracy of the classifier be $P(A)$, then the kappa statistic is defined as:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{6.24}$$

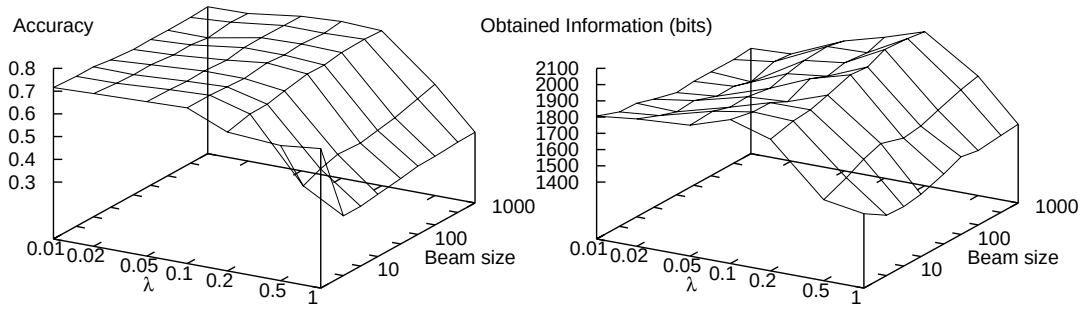
The kappa statistic has a range of -1 to 1, and takes a value of 0 when the performance is no better and no worse than chance.

6.3.6 Parameter estimation

We noted earlier that Equation (6.6) makes some unrealistic independence assumptions. It assumes that the distributional divergence scores between pairs of connectives are logically independent, when in fact they are functions of fixed probability distributions. For example, if there are N connectives then the model assumes that the $\binom{N}{2}$ distributional divergences are independent, when in reality they derive from the N probability distributions. As a result, our model underestimates the value of $P(data|\mathbb{T})$, and so overestimates $L(data|\mathbb{T})$. It is therefore desirable to counterbalance this effect. We do this by weighting $L(data|\mathbb{T})$ by a parameter $\lambda \in (0,1]$, resulting in the following final model:

$$L(\mathbb{T}, data) = L(\mathbb{T}) + \lambda L(data|\mathbb{T}) \tag{6.25}$$

The effect of the parameter λ on the model has a natural interpretation. In general, the use of λ in effect corrects $P(data|\mathbb{T})$ by taking its $\frac{1}{\lambda}$'th root. So, for example, setting λ equal to 0.1 can be interpreted as implying that the original model of $P(data|\mathbb{T})$ assumed ten times as



(a) Accuracy (b) Obtained Information

Figure 6.4: Tuning λ using the validation data

much independence than there actually is, since it corrects the model precisely as if this were the case.

Preliminary experiments were used to tune the λ parameter. To do this, a set of validation data was created in the following way. The set of all ordered pairs of discourse connectives was partitioned into two subsets of equal size. One of these was used for validation. The partitioning had the following property: if the ordered pair of connectives (X,Y) was in the validation set, then its reverse (Y,X) was not, and vice versa. An MDL-based classifier was then constructed, and tested on the validation data with different settings of λ. The data representation of this classifier was based on applying the Kullback-Leibler divergence function to co-occurrences of discourse connectives with other discourse markers. The classifier's performance is summarised in Figure 6.4, which shows that the setting $\lambda = 0.1$ gave the most consistently good results on both the accuracy and Obtained Information metrics. Hence this is the value used in all the results reported below.

Recall that our data is the set of 79 distributional divergence scores between the 79 connectives already in the subtaxonomy and the new connective to be added. Since our original model assumes these 79 scores are independent, a setting of $\lambda = 0.1$ can be interpreted as postulating that the data instead contains $79 \times 0.1 \approx 8$ degrees of independence. Interestingly, from manual analysis of discourse connectives, Knott (1996) reaches the conclusion that the number of independent features required for describing discourse connectives is also 8. The preciseness

of this agreement is not overly important, but these results do show that the two estimates of the number of degrees of independence in the data are at least roughly the same.

6.3.7 Baselines and Upper Bound

To get a feeling for the difficulty of the task, two baseline classifiers were constructed for comparison with the MDL-based model. Since our hypothesis is that the MDL-based model can outperform methods that do not take the global distribution of substitutability relationships into account, these baseline classifiers use only local information. In particular, the baseline classifiers have access to (i) the distributional divergences between pairs of connectives, and (ii) the frequencies of different relationships in the taxonomy before inserting the new connective. However they are not allowed to use the updated frequencies of the different relationships after the new connective has been inserted.

The first baseline classifier used the Naive Bayes technique. Due to the high observed frequency of the EXCLUSIVE relationship, this classifier predicts all pairs to be EXCLUSIVE. This classifier has an accuracy of 69.9%, its amount of Obtained Information was 2280 bits (out of a maximum of 6537 bits), and its Relative Information score was 0.222. The kappa score of this classifier was 0 ($N = 6320, k = 2$). The second baseline classifier was motivated by the principle that if two words have similar distributions then they are likely to also have similar meanings. This classifier assumed that the new connective would be a SYNONYM of the connective X which it was distributionally most similar to. As a result, the new connective's relationship to all other connectives would be identical to X's. We will refer to this classifier as "SYN-with-similar". The confusion matrix of this classifier is shown in Figure 6.2. Its overall accuracy was 70.7%, its amount of obtained information was 3333 bits, its relative information score was 0.307, and its kappa score was 0.292 ($N = 6320, k = 2$). The overall accuracy of the second baseline classifier is not significantly greater than that of the first (using two-tailed $\chi^2, p = 0.34$). The information theoretic measures show a greater difference, due to the second classifier's correct classification of some of the less probable relationships. However one disadvantage of the information theoretic measures is that it is not straightforward to calculate whether differences are significant.

Agreement between human subjects would be an ideal upper bound for the task. The literature on annotation shows that human annotators frequently disagree on classification tasks, and it is reasonable to expect there to also be a degree of disagreement on substitutability. Our experience in studying Knott's taxonomy in detail revealed that there are judgements on sub-

	Predicted relation					Class statistics		
Actual relation	SYN	HYPO	HYPER	CONT	EXCL	Precision	Recall	F-score
SYNONYM	**0**	2	2	12	24	0.00	0.00	0.00
HYPONYM	6	**7**	2	11	26	0.13	0.13	0.13
HYPERNYM	1	0	**3**	11	37	0.04	0.06	0.05
CONT. SUBS.	28	16	18	**723**	971	0.41	0.41	0.48
EXCLUSIVE	100	27	47	509	**3737**	0.78	0.85	0.81

Table 6.2: Confusion matrix for the SYN-with-similar classifier

stitutability made by Knott that we feel are, at the least, questionable. Knott seems well aware of the potential for disagreement, and even advocates that it would have been preferable if his taxonomy had been constructed on the basis of a sizable group of subjects. If such an undertaking has been done, we could then estimate the inter-subject agreement on substitutability between humans, and use this as an upper bound on machine performance. Our experience in manually extending Knott's taxonomy leads us to agree with Knott when he says:

> the amount of data needed in order to build a taxonomy of any reasonable size from scratch makes such an experiment quite infeasible, bearing in mind the huge number of relationships that must be documented. (Knott, 1996, p. 78)

Consider how difficult it would be for a human to perform the task which we are attempting. A subject would have to take a connective and determine its substitutability with about eighty other connectives. In order to do this, many hundreds of texts containing discourse connectives would have to be studied and hundreds of judgements on substitutability would have to be made. Then the whole process would be repeated for every connective in the taxonomy, requiring, in total, thousands or tens of thousand of judgements. This is clearly unrealistic. Indeed, if such a procedure were straightforward then there would be little need to develop an automatic classifier at all. As such, we will only attempt to estimate a reasonable figure for what human agreement on the task might be. To do so, we will make a crucial assumption: that inter-subject agreement on substitutability judgements does not differ significantly from inter-subject agreement on similarity ratings. We consider this the default assumption, in the absence of reasons for believing that there should be a significant difference.

Suppose we have three random variables X_1, X_2 and X_3. The correlations for all three possible pairings of these variables can be calculated; call these r_1, r_2 and r_3. Howell (2002,

pages 280–281) presents a technique (originally due to Williams, 1959) for comparing the three correlation scores. This technique can be used, for example, to determine which of X_1 and X_2 is a better predictor of X_3. Given the three correlation scores, a value of the t distribution is calculated using the following equation:

$$t = (r_1 - r_2)\sqrt{\frac{(N-1)(1+r_3)}{2\left(\frac{N-1}{N-3}\right)\left(1-r_1^2-r_2^2-r_3^2+2r_1r_2r_3\right)+\frac{(r_1+r_2)^2}{4}(1-r_3)^3}} \tag{6.26}$$

The significance of t can then be checked using the standard table. In our case, r_1 will represent inter-subject agreement on substitutability, and r_2 will represent inter-subject agreement on similarity. We will apply this equation in the opposite direction to which it was originally intended. That is, we will assume that t is known, and we will solve for r_1.

Technically, to apply Equation 6.26 we need three random variables defined over the space of pairs of discourse connectives. We let two of these variables be *SIM* and *SUBS*, representing subjects' mean *sim*ilarity ratings and the gold standard *sub*stitutability judgements, respectively. We then assume the existence of a third random variable *INDIVIDUAL*, representing the judgements of an (abstracted) individual subject. We assume the outcomes of *INDIVIDUAL* are some abstract comparisons of pairs of connectives that can be deterministically converted into either similarity ratings or substitutability judgements.

In our case, we know already that inter-subject agreement on similarity ratings is 0.75, i.e. between *INDIVIDUAL* and *SIM* we obtain $r_2 = 0.75$. Furthermore, if we assume the ranking SYNONYMY > HYPONYMY > CONT. SUBS. > EXCLUSIVE, then the correlation r_3 between *SUBS* and *SIM* can be calculated at 0.82 (Spearman's r_s). Both these calculations are based on 48 pairs of connectives, so $N = 48$. The correlation we wish to estimate (r_1) is that between *INDIVIDUAL* and *SUBS*. As we stated above, we will assume there is not a significant difference between r_1 and r_2, so the table for the t distribution says that the maximum allowable value of t is 2.01. Applying Equation 6.26, we deduce that r_1 can be at most 0.854, which we take as an upper limit on inter-subject agreement on substitutability.

We can use this estimated upper-bound on inter-subject agreement to predict upper bounds on the evaluation measures introduced above. To do this, we model the disagreement between subjects by making two assumptions: (i) human subjects only have minor disagreements (e.g. SYNONYMY might be confused with HYPONYMY, but not with EXCLUSIVE), and (ii) subject to this constraint, disagreements are distributed evenly. We implemented these assumptions in a model of disagreement, and experimented with different rates of disagreement between subjects. We found that if two subjects disagree on 13.9% of judgements, then their correlation is

	Accuracy	Obtained Information	Relative Information	kappa
Naive Bayes	69.9%	2280	0.222	0.000
SYN-with-similar	70.7%	3333	0.307	0.292
Upper bound	86.1%	5657	0.685	0.762
Perfect performance	100.0%	6537	1.000	1.000

Table 6.3: Baseline classifiers and an upper bound

the required 0.854. In addition, the same level of disagreement produces a Relative Information score of 0.685, resulting from 5657 bits of Obtained Information, as well as a kappa score of 0.762 ($N = 6320, k = 2$). Note that all these upper bounds are conservative, in that they assume that subjects agree substantially more on substitutability than they do on similarity. The performance of the baseline classifiers and the upper bound is summarised in Table 6.3, however it should not be forgotten that while the baselines are completely empirical, the upper bound is a rational construction based on subjects' similarity ratings.

6.3.8 Results and discussion

We begin by discussing the effects of each of the parameters in turn. We then report some combinations of parameters that gave the best results.

Beam size Some unexpected results emerged when the effects of beam size were examined. Firstly, there was in general no great advantage to searching a larger proportion of the space of possible taxonomies. This suggests that if the correct answer is contained in the beam, then it is likely to occur near the top of the beam during the early stages of the search, and so survive the pruning stages. Secondly, there was actually a minor degradation in performance when large beam sizes were used. A larger beam allows more incorrect answers to be maintained lower in the beam, and it seems that sometimes these leapfrog the better answers towards the final stages of the search, leading to worse results. It is possible that these unexpected results are a consequence of the way we ordered decisions within the beam search. Recall that we first decide on the relationships between the new connective and the connectives to which it is distributionally most similar. If these relationships that are decided first are more likely to be correct, then these can be maintained within a narrow search beam. (In general,

if mistakes are made early by any instantiation of beam search, then a wider beam is required to keep the correct answer "alive" until it hopefully gets promoted within the beam later in the search.) But why should predictions regarding distributionally more similar connectives be more accurate? Consider that there is considerable noise in our co-occurrence data. It may be that lower distributional divergences are in general less noisy than higher divergences. This would follow logically, for example, if there is a general effect of noise producing higher divergences. In any case, the results are welcome. They show that for this task we do not have to use vast computational resources to achieve the best results.

Co-occurrence features To determine whether certain types of co-occurrences led to better results, we analysed the results obtained using simple data representations with different types of co-occurrence features. (Recall that by "simple" we mean that only a single function of the co-occurrence data is used in the model.) Analysis of the mean accuracies (varying other parameters such as beam size) showed that co-occurrences with discourse markers produce better results on all three measures than those with verbs and those with all words (versus verbs: $t = 3.57, df = 196, p < 0.005$; versus all words: $t = 3.50, df = 196, p < 0.005$).

Distributional function Using simple data representations, and varying the type of co-occurrence feature used as well as the beam size, we found that KL divergence gave, on average, greater accuracy than the variation in pointwise entropy function ($t = 3.62, df = 281, p < 0.001$).

Type of data representation In all, 29 different compound data representations were experimented with. As discussed above, these were of three types: Compound$_{cooc}$ used different types of co-occurrence features; Compound$_{func}$ used both distributional functions ($D(X||Y)$ and $V(X,Y)$); and Compound$_{both}$ combined both different co-occurrence features and different functions. Multiple results were then obtained by taking different instantiations of the parameters. These results confirmed that enriching the representation of the lexical co-occurrence data can improve performance. Furthermore, both types of data enrichment (multiple types of co-occurrences, multiple distributional functions) significantly improve performance, both in isolation and in combination. This provides further support for our hypothesis in Chapter 5 that the variation in pointwise entropy function can improve performance on lexical acquisition tasks.

Having discussed the general effects of each of the parameters, we now report on some spe-

Type of data representation	**Parameter settings**			**Performance**			
	Co-occurrences	Functions	Beam	Accuracy	$2 \times I_o$	I_r	kappa
Simple	verbs	$D(X\|\|Y)$	5	74.0%	3836	0.331	0.396
Compound$_{cooc}$	DMs, verbs	$D(X\|\|Y)$	5	73.8%	3820	0.325	0.394
Compound$_{func}$	all words	$D(X\|\|Y)$, $V(X,Y)$	100	**76.2%**	**4158**	0.387	**0.468**
Compound$_{both}$	DMs, verbs, all words	$D(X\|\|Y)$, $V(X,Y)$	9	75.8%	4123	**0.406**	0.446
Naive Bayes	—	—	—	69.9%	2280	0.222	0.000
SYN-with-similar	—	—	—	70.7%	3333	0.307	0.292
Upper bound	—	—	—	86.1%	5657	0.685	0.762

Table 6.4: Best classifiers for different types of data representations

cific instantiations. In particular, we report on the best results achieved using each of the four types of data representations. As we did in Section 6.3.6 when tuning the λ parameter, we partition the set of all ordered pairs of connectives (X,Y) into two equal subsets. One subset is used for validation. The classifiers that were most accurate on the validation data were then evaluated on the second subset, and the results are shown in Table 6.4 (since our test set is halved, we report $2 \times I_o$ to enable easy comparison with the baselines and upper bound). The results show that compound data representations incorporating the function $V(X,Y)$ (sub-types *func* and *both*) can achieve over 75% accuracy, whereas no classifier that did not use $V(X,Y)$ performed above this level. Classifiers using compound data representations incorporating $V(X,Y)$ also achieved higher on three other metrics than any classifier not using $V(X,Y)$ (the kappa scores reported in Table 6.4 are with $N = 3160, k = 2$). This provides further demonstration of the utility of the new function for distributional analysis, and shows that combining complementary distributional functions can yield better results than combining different types of lexical co-occurrences. The confusion matrix and analyses per class for the best performing Compound$_{func}$ classifier are shown in Table 6.5.

	Predicted relation					Class statistics		
Actual relation	SYN	HYPO	HYPER	CONT	EXCL	Precision	Recall	F-score
SYNONYM	**0**	2	0	5	13	0.00	0.00	0.00
HYPONYM	0	**7**	0	0	22	0.12	0.24	0.16
HYPERNYM	0	0	**4**	5	14	0.10	0.17	0.13
CONT. SUBS.	4	21	13	**575**	265	0.52	0.65	0.58
EXCLUSIVE	5	27	25	330	**1823**	0.85	0.82	0.83

Table 6.5: Confusion matrix for the best performing classifier

6.4 Experiment 11: Developing ensembles for computer-assisted taxonomy development

6.4.1 Introduction

As discussed earlier, the combining of classifiers via ensemble methods has two advantages over the use of individual classifiers. Firstly, the ensembles may improve overall performance on a task. Secondly, the outcome of the voting can be treated as a type of confidence score. In semi-automatic applications, this may guide the human user towards automatic classifications that were based on borderline decisions and are more likely to be doubtful (Osborne and Baldridge, 2004).

However ensemble methods are restricted in their practical applications. In particular, restrictions arise from the implicit assumption that the classification of each item is independent of all other items. This means ensemble methods cannot always be applied when there are global constraints on classification. In parsing, for example, ensembles can vote on parse constituents, but there is no guarantee that the set of winning constituents form a parse tree free of crossing brackets. Henderson and Brill (1999) address this problem, by proving that the set of selected constituents can be guaranteed to contain no crossing brackets if each constituent is agreed upon by more than half of the individual parsers. Similar problems arise in the task of automatic taxonomy construction, as can be seen from the following example. Suppose we have three connectives A, B and C, and we have three classifiers for predicting substitutability relationships between them. Figure 6.5 shows that the ensemble of the three classifiers need not produce a consistent set of relationships, even when over half the classifiers agree on each

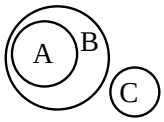
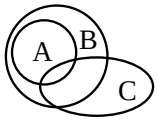
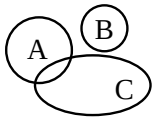
	Pairwise relationships	**Consistent?**	**Venn diagram**
Classifier 1	HYPONYM(A,B), EXCLUSIVE(A,C), EXCLUSIVE(B,C)	Yes	
Classifier 2	HYPONYM(A,B), CONT. SUBS.(A,C), CONT. SUBS.(B,C)	Yes	
Classifier 3	EXCLUSIVE(A,B), CONT. SUBS.(A,C), EXCLUSIVE(B,C)	Yes	
Ensemble	HYPONYM(A,B), CONT. SUBS.(A,C), EXCLUSIVE(B,C)	No	*Not possible*

Figure 6.5: Consistent classifiers leading to an inconsistent ensemble

pairwise relationship (cf. Henderson and Brill, 1999). It follows that we cannot apply ensemble methods directly to the task of extending or constructing a taxonomy. Instead, in this final experiment of the thesis, we apply ensemble methods to predicting pairwise substitutability relationships, with the intention that these predictions might be of assistance within the context of computer-assisted taxonomy development.

6.4.2 Hypotheses

The aims of this experiment are twofold. The first objective is to improve the performance in predicting substitutability between pairs of connectives. For an ensemble of classifiers to outperform its components, the errors made by the individual classifiers must be sufficiently varied.

Hypothesis 6.2 *Ensemble methods can improve the performance in predicting substitutability relationships.*

The second objective is to determine whether ensembles might provide additional useful information for the semi-automatic development of taxonomies of discourse connectives.

Hypothesis 6.3 *The voting results of ensembles can be interpreted as confidence scores. That is, the greater the number of votes a particular classification receives, the more likely it is to be correct.*

6.4.3 Methodology

Ensembles were constructed using MDL-based classifiers from the previous experiment. We did not create all possible ensembles. Instead, we reduced the combinatorial possibilities for creating ensembles in two ways. Firstly, only classifiers with a beam size of 5 were used, as analysis of the results of the previous experiment showed that larger beam sizes beyond this did not give consistently better results. Secondly, we supposed that someone performing lexical acquisition experiments would be interested in maximising one of the four evaluation measures introduced in the previous experiment. These were accuracy, Obtained Information, Relative Information, and kappa. Therefore, for each of the three evaluation measures in turn, we compiled a list of the 20 classifiers which performed best on this measure. The lists of classifiers used can be found in Appendix E. This selection was done using the validation data described in the previous experiment; the results reported below are on the separate test data. From each of the lists of 20 classifiers, we constructed ensembles using the top N classifiers, for $N = 1, \ldots, 20$. For each evaluation measure, 80% of the classifiers used complex data representations, while 20% used simple data representations. On average, 58% incorporated the variation in pointwise entropy function, and 73% used co-occurrences with discourse markers.

6.4.4 Results and discussion

Figure 6.6 plots the performance of the ensembles against the number of individual classifiers they contain. On three of the four measures, performance initially improves slightly as the size of the ensemble increases, peaking at around $N = 13$, with an accuracy of 80.0%. This accuracy is significantly higher than the best performing individual classifier from the previous experiment (accuracy=76.2%) ($\chi^2 = 16.0, df = 1, p < 0.0001$), and as a result it is also higher than the baseline. A new high is also achieved for kappa: $\kappa = 0.510$ ($N = 3160, k = 2$). However the amount of Obtained Information is more volatile.[5] In shifting from $N = 1$ to $N = 2$, it plunges by over 210 bits. This indicates that the ensemble of 2 classifiers is predicting less likely relationships (e.g. HYPONYMY) less often. This is a direct consequence of the ensemble

[5]Ensemble voting sometimes results in a tie, in which case we split the prediction between the joint victors. In such cases, we use the generalised definitions of Obtained Information and Relative Information given by Kononenko and Bratko (1991).

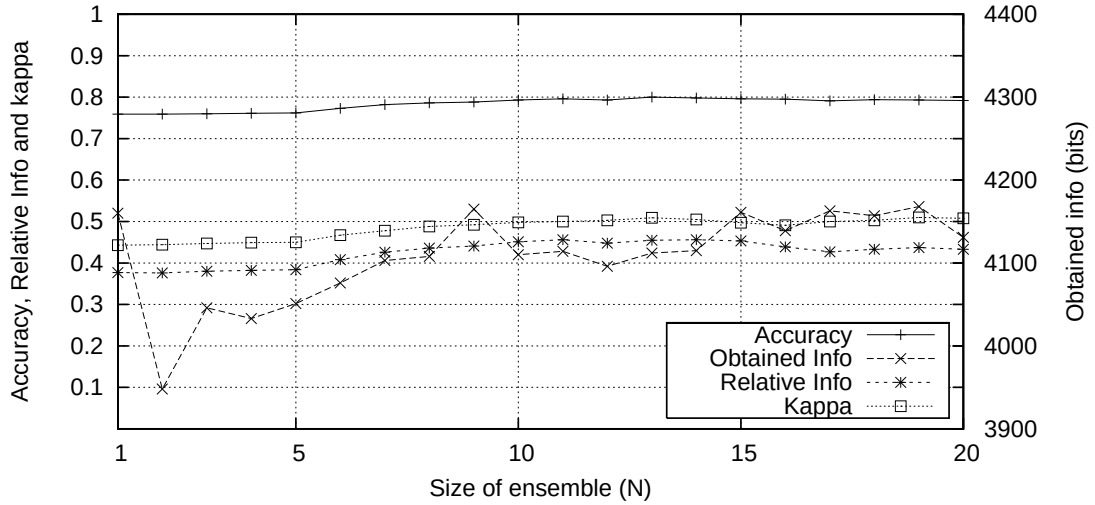


Figure 6.6: Ensemble results

voting process. Because all votes are given equal weight, there is effectively a bias towards the higher frequency classes. To overcome this bias, a voting system based on information theory was also used. Suppose a classifier predicts relationship *rel* for a pair of connectives. The greater the prior likelihood of *rel*, the lower the weighting of this vote in the ensemble. In particular, it is given a weight of $-\log_2 P(rel)$. Figure 6.7 shows that this voting system improves performance on the I_o evaluation measure. The best result by an ensemble using the information theoretic voting system is 4358 bits. This is 66.6% of the number of bits possible, and an improvement of about 5% over the best achieved by any individual classifier (4158 bits). There is thus evidence that performance on all three measures can be improved by using ensemble methods, supporting our Hypothesis. However, we have not found that a single ensemble method can improve performance on all evaluation metrics. Instead, the choice of ensemble method is dependent on which evaluation metric the experimenter wishes to optimise.

The results obtained using ensembles are compared with the best classifiers based on simple and complex data representations in Figure 6.8. It shows that there is a general improvement in performance as the complexity of the machine learning techniques increases.

The likelihood of an ensemble's prediction being correct was also compared with the number of votes it received. To do this, the ensemble of 20 classifiers used for optimising accuracy was used (so the standard voting system is used). Figure 6.9 shows that as the num-

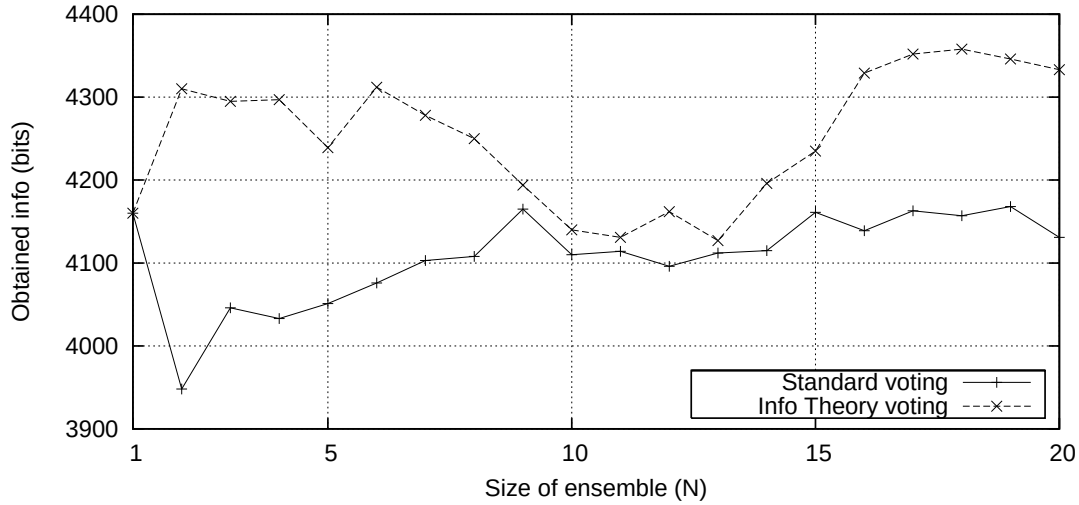


Figure 6.7: Ensemble results using different voting methods

ber of votes received increases, so too does accuracy (Pearson's product-moment correlation $= 0.95, p < 0.001$). So if the ensemble were to be used in assisting semi-automated taxonomy construction, then the voting results could help the human find mistaken predictions. This provides support for Hypothesis 6.3. If all predictions receiving less than a given number of votes are rejected, then the rejection threshold determines a tradeoff between precision and recall. For example, precision in predicting pairwise relationships of over 90% can be obtained if it is acceptable for recall of those relationships to drop below 50%.

6.5 Summary

In this chapter we introduced a method for modelling the global structure of the lexicon. A Minimum Description Length model was defined over taxonomies of discourse connectives, so that both the prior likelihood of the taxonomy, and how well it accounts for the data are taken into account. The prior uses a multinomial term to assign the highest probability to the taxonomy which contains relationships in proportion to their prior likelihoods. The posterior models how well the taxonomy explains lexical co-occurrence data. However the modelling is at one level of abstraction: the *data* is taken to be distributional functions of the raw co-occurrences. Both the prior and the posterior are expressed in terms of products over all pairwise relationships. This requires some independence assumptions (which we correct for using a parameter λ), but

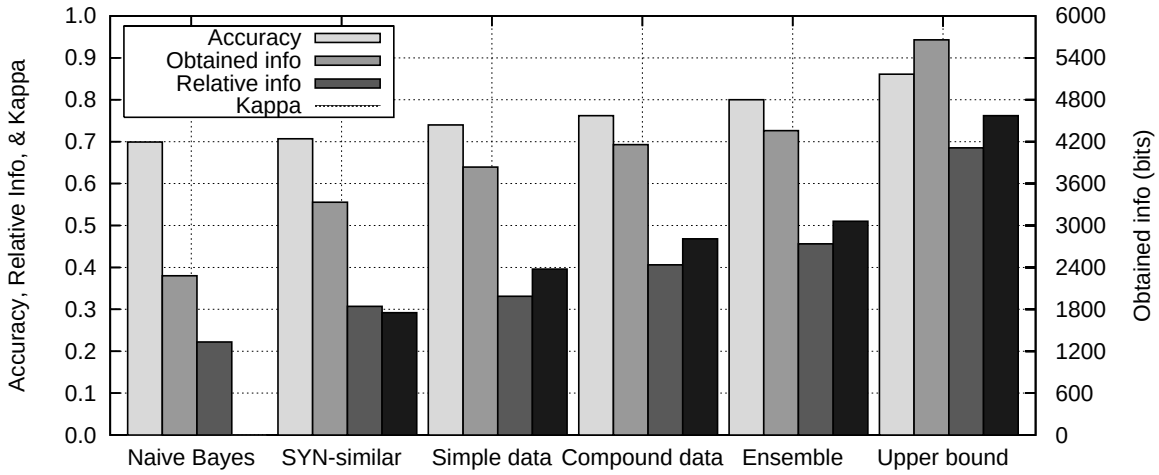


Figure 6.8: Comparison of different classifiers with baselines and the upper bound

it allows both a mathematically elegant formulation and an efficient implementation of beam search using dynamic programming techniques.

Experiments with the MDL-based model showed that it outperformed two different baseline classifiers. The increases in accuracy over the baselines demonstrated that the challenges imposed by the high prior likelihood of EXCLUSIVE can be overcome, and up to 80.0% accuracy was achieved. However, the improvement over the baselines was most striking in the increase in the amount of Obtained Information, which can be considered a measure of how much knowledge the classifier has successfully learnt. Up to 66.6% of the bits of information in the taxonomy were learnt by an ensemble, compared to 53.4% and 36.6% by the two baselines. We also found further evidence for the utility of the variation in pointwise entropy function introduced in Chapter 5. Although on its own it was not as useful as Kullback-Leibler divergence, in combination they achieved significantly better results than KL divergence in isolation. We showed that ensembles cannot be guaranteed to produce a consistent set of predictions, even when Henderson and Brill's "majority of votes" condition holds. Nevertheless, experiments with ensemble methods show that they can improve performance on predicting pairwise relationships, and that the voting results can be treated as confidence scores accompanying the predictions.

Figure 6.8 demonstrates that there is still a great divide between the results obtained and our estimated upper bound for the task. There is the potential for many further advances to be

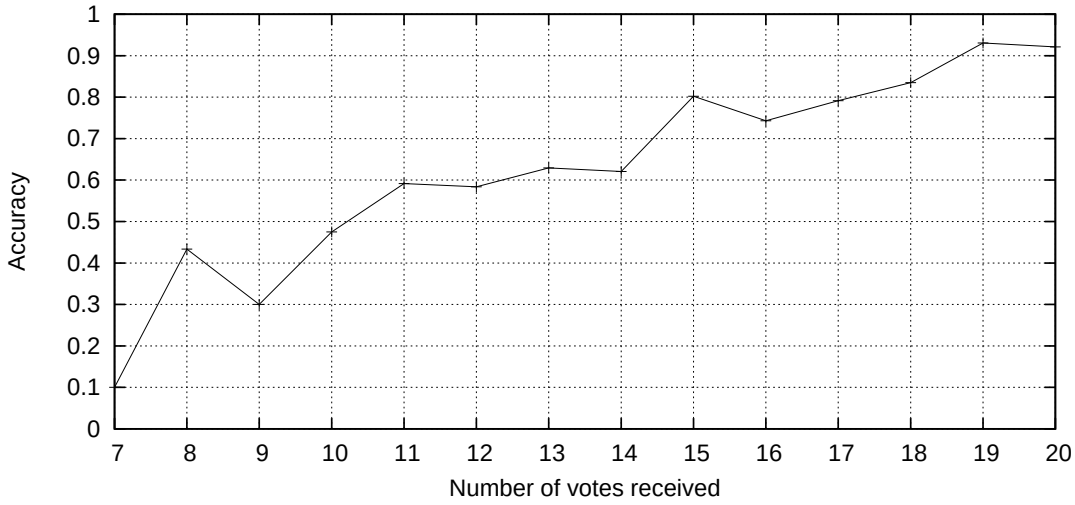


Figure 6.9: Analysis of predictions made by an ensemble

made. In particular, the differences in the Obtained Information metric show that the less likely substitutability relationships currently have a lower recall than the more likely ones. However it is worth considering that the results we have achieved are in the face of (i) sources of noise in our data, (ii) processing of data resulting in a net information loss, and (iii) unrealistic theoretical assumptions, for example of independence.

Chapter 7

Conclusions

This chapter summarises the major contributions of the thesis. It also outlines directions for future work.

7.1 Summary of contributions

This thesis constitutes the first broad coverage study of the automatic acquisition of knowledge about discourse connectives. Previous chapters have considered both the classification of individual connectives, as well as the learning of relationships that hold between connectives. In the process, four major contributions were made.

Firstly, this thesis has demonstrated that semantic information about discourse connectives can be acquired automatically from unannotated resources, despite a degree of noise in the data. This has previously been demonstrated for word classes such as nouns and verbs, but not for words involved in signalling discourse relations. This thesis showed that automatic web mining and corpus analysis techniques can be applied with a sufficiently low degree of error to enable the statistical study of distributional properties of discourse connectives. A major requirement for this was that discourse connectives could be positively identified with a sufficiently high degree of accuracy. Identifying discourse connectives is made difficult by the high level of ambiguity of many phrases which only sometimes function as discourse connectives. Because we are interested in identifying the clauses which discourse connectives relate, an automatic statistical parser was applied in order to detect clause boundaries, including those of embedded clauses. The parse trees were analysed automatically to detect the presence of discourse connectives at the clause boundaries. A manual evaluation revealed this process of automatically identifying discourse connectives has an error rate of about 12%. However the

question of whether this level of noise is permissible must ultimately be answered empirically. The experiments throughout this thesis testify that many significant results can be obtained when the data contains this level of noise.

The world wide web has huge potential as a source of linguistic data. Techniques for utilising this potential were developed in order to obtain statistical data on the distributions of less common discourse connectives. These techniques involved automatically conducting web searches for the surface forms of discourse connectives. An off-the-shelf HTML parser was used to extract textual elements from web pages, and these textual elements were automatically analysed for discourse connectives. Analysis of discourse marker bigrams obtained from the web and from the British National Corpus showed a high degree of correlation between the two sources of data.

Secondly, this thesis has demonstrated that the semantic similarity of pairs of discourse connectives correlates with their distributional similarity, where distributional similarity is measured via lexical co-occurrences. This extends previous results obtained for other classes of words, such as nouns and verbs (e.g. Miller and Charles, 1991; Resnik and Diab, 2000). However discourse connectives are sensitive to a wide range of deep semantic and pragmatic properties of texts, and it is therefore not obvious that lexical co-occurrence distributions should capture semantic similarities. The most informative kind of co-occurrence was found to be that of multiple discourse markers within a sentence. However it was also found that co-occurrences with verbs produced significant results when classifying discourse connectives according to their veridicality.

The distributional similarity of pairs of discourse connectives was also found to be related to their substitutability. In particular, pairs of discourse connectives that are always substitutable have the most similar distributions, whereas connectives which are never substitutable have the least similar distributions. In addition, if two connectives stand in a HYPONYMY relation, then the co-occurrence distribution of the HYPERNYM is likely to have a greater entropy then that of the HYPONYM.

The third contribution of this thesis was the introduction of a new function V for comparing distributions of lexical items. Unlike previous functions, this new one does not attempt to measure distributional similarity. Instead, V measures the variance in surprise in seeing one lexical item in place of another. Experiments showed that V is sensitive to the substitutability of discourse connectives. This is most clear in the case of HYPONYMY, for which V both takes the highest values in general, and shows the greatest sensitivity to the order of its arguments. The new function V was also found to be of practical utility in lexical acquisition tasks. Models

that combined both V and distributional similarity were better at predicting substitutability relationships than models based solely on similarity.

Finally, the fourth contribution of this thesis was to introduce a technique for modelling the global structure of the lexicon, based upon the Minimum Description Length principle. To do this, a lexical taxonomy was represented as a set of pairwise relationships, and a multinomial prior was defined over this set. The posterior probability was estimated by considering how well each pairwise relationship predicted the distributional similarity of the two connectives. The factorisation of both the prior and posterior probabilities into products over all pairwise relationships enables dynamic programming techniques to be applied in the search for the optimal taxonomy. The new model was more accurate at predicting substitutability relationships than simpler techniques which do not consider global aspects of the lexicon.

The contributions described above were demonstrated through a range of machine learning experiments of three distinct types: acquiring properties of individual connectives, learning relationships between pairs of connectives, and constructing taxonomies containing multiple connectives. These experiments were presented as complementary, however the motivations behind them are inter-related, as are the conclusions that can be drawn from them. The experiments into learning semantic properties (Chapter 4) were based on the hypothesis that words sharing a semantic property also have similar distributions. As such, the experiments are somewhat like discrete versions of tests of the distributional hypothesis, since distributional similarity is compared with the discrete category judgements of the gold standards. However the support for the distributional hypothesis is not strictly required for discrete classification. This is partly because correlation calculations are affected by the presence of heterogeneous subsamples (Howell, 2002), e.g. if different subclasses of discourse connectives show different empirical trends, this may make it hard to find a significant overall correlation. That we also found support for the distributional hypothesis (Chapter 5) thus constitutes a stronger finding.

The relationship between taxonomy construction (Chapter 6) and learning relationships between pairs of discourse connectives is more straightforward. This is because in our experiments taxonomies were just compact representations of pairwise substitutability relationships, and we also tried to predict these relationships in isolation. Substitutability is inherently a more complex notion than similarity, for example a single similarity score cannot be used to distinguish the two possible directions of the HYPONYMY relationship. However we found that similarity is affected by the substitutability of connectives (Chapter 5). Similarity relationships between multiple discourse connectives can be captured in a single data representation, e.g. a hierarchical clustering of connectives. To summarise, this thesis has provided the tools nec-

essary for automatically creating a taxonomy or clustering of connectives and also labelling the connectives in the taxonomy/clustering with semantic properties. Knott (1996) has argued that empirically-motivated taxonomies can be used as evidence for sets of semantic properties which are consistent with the taxonomy. However the task of deciding which particular set of semantic properties is best supported by empirical data has been beyond the scope of this thesis. In principle, though, both automatically constructed clusters and automatically constructed taxonomies can be inspected manually in an attempt to discover semantic properties of connectives that have previously eluded intuitions.

In summary, this thesis has made significant contributions to the field of automatic lexical acquisition. These include both technical advances, as well as the broad coverage application of machine learning techniques to acquiring knowledge about discourse connectives. The latter was made possible by developing new automatic techniques for processing discourse connectives.

7.2 Brave new words: Studying new connectives

Imagine that a new English discourse connective has been discovered in an isolated community which speaks a unique dialect of English. Let us suppose the new subordinator is *wockerjabby*, and that fortunately the written records of this community provide large amounts of empirical data on how it is used. How could an analyst use the techniques developed in this thesis to analyse its meaning?

Several methods of proceeding are possible. However a sensible first step would be to determine which other connectives *wockerjabby* is distributionally similar to. In doing so, lexical co-occurrences would be collected for use in the distributional representation, using the automated methods described in Chapter 3. To do this, it would be necessary to modify the parser so that it recognises *wockerjabby* as a subordinating conjunction. We would choose discourse marker co-occurrences for the distributional representation because similarity of these co-occurrences is known to correlate with semantic similarity (Chapter 5). This will enable us to make an educated guess as to which other connectives are semantically similar to this one.

However semantic similarity is a one dimensional property, and we might next want to clarify precisely how it is related to these similar connectives. For example, *because* and *so* are similar in many respects (both are veridical and signal causal relations), but not in others (they signal opposite directions of causation). If it were the case, however, that we could predict that *wockerjabby* is a SYNONYM of another connective (which has already been studied), then

this fact would tell us everything we need to know about *wockerjabby*. How can we determine if *wockerjabby* has a SYNONYM? We found that the best way to predict substitutability relationships was through the insertion of a connective into a taxonomy (Chapter 6). Furthermore, predicting substitutability has higher accuracy when the posterior model combines both distributional similarity and the variance in pointwise entropy function (V). So we proceed by calculating values of V, incorporating these in the posterior model, and using beam search to insert *wockerjabby* into our taxonomy. If the model does predict that *wockerjabby* has a SYNONYM, then our work is done.

However SYNONYMY is quite rare, so let us suppose that our model does not predict any SYNONYMS for *wockerjabby*. In fact the most likely situation is that *wockerjabby* will be either CONTINGENTLY SUBSTITUTABLE or EXCLUSIVE with the majority of connectives, perhaps being related by HYPONYMY to just one or two. So let us suppose the model predicts such a situation. The most informative relationships will be those where *wockerjabby* is either EXCLUSIVE with, or a HYPONYM of, other connectives. In such cases, monotonic inferences can be made. For example, if *wockerjabby* is EXCLUSIVE with *if* then we can conclude that the clause which is the complement to *wockerjabby* cannot be the antecedent of a conditional relation.

It is likely, however, that the substitutability relationships between *wockerjabby* and other connectives would not fully determine the semantics of *wockerjabby*. For example, few if any useful inferences can be made from CONTINGENTLY SUBSTITUTABLE relationships. A more direct approach to learning semantic properties of *wockerjabby* might therefore be required. A series of machine learning experiments could therefore be performed to determine whether *wockerjabby* signals a VERIDICAL relation, whether it signals a relation with NEGATIVE POLARITY, and so on (Chapter 4). It would be particularly interesting if *wockerjabby* had a combination of semantic properties that no other connective exhibits. What could such a combination be? One combination that is conspicuously absent from the taxonomy we described in Chapter 3 is the following: **veridicality**=NON-VERIDICAL, **type**=CAUSAL, **direction**=BACKWARD. A discourse connective with these properties would essentially be very similar to *if*, except that it would take as its complement not the antecedent clause but the consequent clause. So if *wockerjabby* had these properties then the following four sentences would be paraphrases:

(7.1) *If* you're tired, go to bed.

(7.2) Go to bed *if* you're tired.

(7.3) You're tired *wockerjabby* go to bed.

(7.4) *Wockerjabby* go to bed, you're tired.

To summarise, the techniques developed throughout this thesis can in principle be applied to the study of new discourse connectives. Furthermore, we have suggested a procedure that an analyst could follow: first predicting similarity, then substitutability, then semantic properties. The information gained from these different experiments would be somewhat complementary, but not entirely so, for example if two connectives have different semantic properties then this has consequences for their substitutability. As such, the predictions of each experiment could be compared to check whether the different types of knowledge about discourse connectives are converging. If so, we could be confident of arriving at a sensible analysis of the new connective.

7.3 Complications and simplifications

The human language interpreter is remarkably robust when confronted with noisy data. Computers are less so. The accomplishments of the thesis should be seen in the context of the various sources of noise and bias in the data, and also of various assumptions we made in developing our models. We now elaborate on each of these in turn.

Firstly, many of the algorithms that we employed introduced a degree of noise into our data. We use an automatic method for identifying discourse connectives, however we estimate the error rate of this method to be about 12%. This method was also designed to sacrifice recall for the sake of precision, on the basis that the web can provide, in practice, as much data as is required. However it is possible that this emphasis on precision led to our collecting a skewed sample of connectives. Our use of the web as a source of linguistic data may also have led to a skewed sampling of certain usages of discourse markers, as the web is not a balanced corpus. We employed a statistical parser in order to detect clause boundaries, and to provide a source of word class information. However the parser has an error rate of about 10%, leading to further noise in our data.

Secondly, we made certain modelling assumptions that resulted in our ignoring certain aspects of the data. When predicting substitutability relationships, we employed two functions for comparing co-occurrence distributions (distributional divergence and variation in pointwise entropy), however these functions ignore certain aspects of the co-occurrence data. For example, frequency information is lost, despite this potentially being a useful statistic (for example, we found it to be useful for distinguishing the order of HYPONYMY). We also smooth our

distributional functions to avoid problems arising from zero counts, however no attempt was made to optimise the degree of smoothing. In our MDL-based model, we make some key independence assumptions in our equations for the posterior probability. These are in general not justified, however we are able to correct for this somewhat by weighting $L(data|\mathbb{T})$ by a parameter λ. Finally, we are not able to search the space of taxonomies exhaustively, in order to find the global optimum. Although we found that increasing the beam size from 10 to 1000 caused a minor degradation in performance, it is possible that a complete search might give better results.

7.4 Directions for future work

This thesis has introduced new techniques, and has established empirical results for a subset of English discourse markers. We conclude by outlining what we consider to be the most important directions for future research.

Empirical study of substitutability judgements One line of future work would be an empirical study of the degree to which speakers agree on the substitutability of discourse connectives. This thesis used a taxonomy of discourse connectives as a source of gold standard judgements on substitutability. This taxonomy was constructed on the basis of judgements by only two speakers (Alistair Knott and the current author), ignoring some important facts. Firstly, humans do not always agree on linguistic judgements. Secondly, humans sometimes have trouble making categorical judgements. Instead they may feel that some cases are borderline. Thirdly, humans can be a richer source of data than categorical judgements allow for. For example, they may feel that two examples are both bad, but that one is much worse than the other. Further studies are required to determine how these factors relate to judgements on the substitutability of discourse connectives. It is likely that humans will sometimes disagree on specific instances of substitutability. In Chapter 6, we estimated a bound on their level of disagreement based on inter-subject agreement on similarity, but ideally this would be determined empirically. As discussed in Chapter 3, if substitutability is to be studied in a scientifically rigorous manner, it is also necessary to formalise the conditions under which a set of judgements is taken to be sufficient evidence for a relationship such as HYPONYMY.

Co-occurrences of discourse markers Constraints on the interpretation of co-occurring discourse markers is another area requiring further study. This thesis has identified that the

tendencies of discourse markers to co-occur in certain patterns can be informative. In Chapter 3 we discussed some of the hard constraints that are imposed on such co-occurrences, for example inconsistent discourse markers cannot take the same arguments. There are three lines of research in this area that could be pursued further. Firstly, the interpretation of discourse adverbials following discourse connectives is still poorly understood, despite some work by Webber et al. (2003). Secondly, constraints on discourse marker co-occurrences might be used to develop a methodology for exploring the semantics of discourse connectives. Thirdly, statistical tendencies in the co-occurrence of discourse markers remain to be explored and explained. For example, preliminary analysis shows that that TEMPORAL discourse adverbials are more likely to co-occur near TEMPORAL discourse adverbials, while CAUSAL discourse connectives are often syntactically subordinate to other CAUSAL connectives (Hutchinson, 2004c).

Improved identification of discourse markers Our algorithm for identifying discourse markers automatically was sufficiently accurate for many significant results to be obtained. However greater accuracy might lead to support for the hypotheses for which our results did not quite provide significant evidence (e.g. some of our hypotheses concerning the new variation in entropy function might be supported if the data contains less noise). Discourse parsing tasks would also benefit from better identification of discourse markers. One possible avenue of exploration is the use of co-occurrence statistics to aid discourse marker identification. For example, if identifying an instance of a word or phrase as a discourse connective would result in it having highly unusual lexical co-occurrences, then such an instance might be rejected. However there is also the danger that an approach such as this would lead to skewed sampling of the data.

Application to further discourse markers The techniques developed in the thesis could be extended to cover a wider range of discourse markers. Acquiring information about certain types of discourse markers was beyond the scope of this thesis. Firstly, we deliberately excluded polysemous discourse connectives from experiments where this ambiguity would prove problematic. However polysemy is an important issue in Natural Language Processing. Further experiments are required to determine what effects polysemy has on co-occurrence distributions, and whether the polysemy of discourse connectives might be predicted automatically. Secondly, another class of discourse markers that was excluded from the experiments was discourse adverbials. These present difficulties because the discourse coherence relations that they signal have one argument that must be resolved anaphorically. A first step towards

automatically acquiring information about discourse adverbials might be to use co-occurrence distributions based solely on co-occurrences within the clause in which the adverbial appears. Once the limitations of this restricted approach have been explored, heuristics for identifying anaphoric arguments to coherence relations might be incorporated to see if this improves classification accuracy. Thirdly, this thesis has also restricted its scope to English connectives. For other languages, it may be that the various subclasses of lexical co-occurrence are more or less informative for lexical acquisition tasks. It remains to be explored how the techniques we have developed can be applied to languages for which taxonomies such as Knott's are not available.

Further study of variance in surprise The function V that measures variance in surprise is expected to be useful for lexical acquisition tasks involving other classes of lexical items. Since V was shown to be highly sensitive to HYPONYMY, a promising avenue for further work is to study whether similar effects can be found for hyponymy between nouns, using a resource such as WordNet. Studies in automatic thesaurus extraction tend to produce ranked lists of neighbours for each noun. These lists tend to contain both near-synonyms as well as hyponyms, antonyms and co-hyponyms. If the function V is found to be sensitive to noun hyponymy, for example, it might be used re-order such lists in order to predict finer-grained semantic distinctions.

Processing of discourse marker tokens This thesis has involved developing distributional representations for discourse markers. As well as being of use for lexical acquisition purposes, such representations could also be applied to discourse processing tasks. One problem that arises in discourse parsing is that a sentence containing two discourse connectives, for example S_1 *and* S_2 *or* S_3, contains a structural ambiguity. This ambiguity might be resolved by considering the likelihood of the various lexical co-occurrences that are entailed by the two alternative bracketings. In this example, one could compare the likelihood of the words in S_1 occurring in the left coordinate of *or* with the likelihood of the words in S_3 occurring in the right coordinate of *and*. There are potential applications in text paraphrasing too. Suppose we wish to generate a paraphrase for a text by replacing one discourse marker with another. There might be several appropriate candidate discourse markers, and the choice between them might be made on the basis of the lexical co-occurrences that would result.

Appendix A

The gold standard taxonomy

This appendix summarises the extensions to Knott's (1996) taxonomy that were made to produce the gold standard for the experiments reported in Chapters 5 and 6. It also documents the correction of some minor inconsistencies in Knott's taxonomy, which were resolved following personal communication with Knott.

A.1 Resolving inconsistencies in Knott's taxonomy

The part of Knott's taxonomy containing Negative Polarity discourse markers (page 186) has some inconsistencies. Specifically:

1. The discourse markers *while (sense 2), whereas, though, although* and *even though* are all entered in two distinct locations. Furthermore, for each of these markers an EXCLUSIVE relation can be deduced between its two entries.

2. The discourse markers *but, yet* and *however* are all represented as EXCLUSIVE to each of *although, though* and *even though*. This seems wrong, and also contradicts an earlier version of the taxonomy presented by Knott and Dale (1994).

3. The discourse markers *but, yet* and *however* are all represented as EXCLUSIVE to *and*, however it is possible to use *and* in situations where there is a contrast relation that the writer does not wish to signal explicitly.

The corrections proposed by Knott were:

1. For each of the discourse markers with multiple entries, delete the entry near the left hand edge of the diagram (close to the entry for *either*(2)).

2. Delete the leftmost edge entering the top of the node containing *although, though* and *even though*. This makes them CONTINGENTLY SUBSTITUTABLE with *but, yet* and *however*.

3. Make the node that immediately dominates *and* and *otherwise* CONTINGENTLY SUBSTITUTABLE with the node containing *but, yet* and *however*.

A.2 Extending Knott's taxonomy

Figure A.1 provide an alphabetical listing of all discourse connectives in the expanded taxonomy. Figures A.2–A.13 contain fragments of the extended taxonomy that relate connectives absent from Knott's (1996) taxonomy to ones already in it. Each fragment is referred to by a connective it contains that was missing from Knott's (1996) taxonomy. The format of each fragment is a list of substitutability relationships and their arguments. From the relationship of a new connective to one previously in Knott's taxonomy, then additional relationships can be inferred. For example, the "*only when* fragment" states that *when* is EXCLUSIVE with *except when*. Since Knott's original taxonomy states that *after* is a HYPONYM of *when*, it follows that *after* is also EXCLUSIVE with *except when*.

The notation used for specifying the pairwise relationships is that introduced in Chapter 6. We use $\langle rel, X, Y \rangle$ to denote that connectives X and Y are in relationship *rel*, where *rel* is one of SYNONYM, EXCLUSIVE, HYPONYM, or HYPERNYM or CONTINGENTLY SUBSTITUTABLE. If two connectives are in the same fragment of the taxonomy, but no relationship is specified between them, then CONTINGENTLY SUBSTITUTABLE should be inferred. The HYPONYMY relationship is asymmetric; $\langle \text{HYPERNYM}, X, Y \rangle$ should be interpreted as "X is a HYPERNYM of Y", or equivalently "Y is a HYPONYM of X".

after, although, and, as [sic], as(1), as(2), as(3), as long as, as soon as, assuming that, because, before, but, but not after, but not when, but then, by the time, considering that, despite the fact that, else, even after, even before, even if, even though, even though, even when, ever since, except, except after, except before, except since, except when, for, for fear that, for the reason that, given that, however, if, if ever, if only, in case, in order that, insofar as, in that, in the hope that, just as, lest, much as, notwithstanding that, now, now that, once, on condition that, only after, only before, only if, only until, only when, on the assumption that, on the grounds that, or, or else, or rather, presumably because, provided that, seeing as, since, so, so that, supposing that, the instant, the moment, then(1), the way, though, to the extent that, unless, until, until after, when, whereas, whether or not, which is why, which was why, while(1), while(2), yet,

Figure A.1: Discourse connectives in the expanded taxonomy. Parentheses indicate Knott's (1996) sense numbers. The connective *as* occurs in Knott's taxonomy both with and without sense numbers.

⟨HYPERNYM,"until","only until"⟩
⟨HYPERNYM,"until","until after"⟩
⟨HYPERNYM,"except after","except since"⟩
⟨EXCLUSIVE,"except after","only until"⟩
⟨EXCLUSIVE,"except since","only until"⟩
⟨EXCLUSIVE,"only until","until after"⟩
⟨EXCLUSIVE,"until","by the time"⟩
⟨EXCLUSIVE,"only until","by the time"⟩
⟨EXCLUSIVE,"until after","by the time"⟩

Figure A.2: The *only until* fragment

⟨SYNONYM,"even though","despite the fact that"⟩
⟨HYPERNYM,"despite the fact that","notwithstanding that"⟩
⟨HYPERNYM,"despite the fact that","much as"⟩
⟨HYPERNYM,"even though","notwithstanding that"⟩
⟨HYPERNYM,"even though","much as"⟩

Figure A.3: The *notwithstanding that* fragment

⟨HYPERNYM,“so”,“which is why”⟩
⟨HYPERNYM,“so”,“which was why”⟩
⟨EXCLUSIVE,“which is why”,“which was why”⟩

Figure A.4: The *which is why* fragment

⟨HYPERNYM,“in case”,“lest”⟩
⟨EXCLUSIVE,“unless”,“in case”⟩
⟨EXCLUSIVE,“unless”,“lest”⟩
⟨EXCLUSIVE,“unless”,“for fear that”⟩

Figure A.5: The *in case* fragment

⟨SYNONYM,“else”,“or else”⟩
⟨HYPERNYM,“or”,“and/or”⟩

Figure A.6: The *else* fragment

⟨HYPERNYM,“but”,“but then”⟩

Figure A.7: The *but then* fragment

⟨HYPERNYM,"only when","only after"⟩
⟨HYPERNYM,"even when","even after"⟩
⟨HYPERNYM,"when","only after"⟩
⟨EXCLUSIVE,"only when","even when"⟩
⟨EXCLUSIVE,"only when","even after"⟩
⟨EXCLUSIVE,"only when","but not when"⟩
⟨EXCLUSIVE,"only when","but not after"⟩
⟨EXCLUSIVE,"only when","except when"⟩
⟨EXCLUSIVE,"only when","except after"⟩
⟨EXCLUSIVE,"only after","even when"⟩
⟨EXCLUSIVE,"only after","even after"⟩
⟨EXCLUSIVE,"only after","but not when"⟩
⟨EXCLUSIVE,"only after","but not after"⟩
⟨EXCLUSIVE,"only after","except when"⟩
⟨EXCLUSIVE,"only after","except after"⟩
⟨EXCLUSIVE,"even after","but not when"⟩
⟨EXCLUSIVE,"even after","but not after"⟩
⟨EXCLUSIVE,"even after","except when"⟩
⟨EXCLUSIVE,"even after","except after"⟩
⟨EXCLUSIVE,"even when","but not when"⟩
⟨EXCLUSIVE,"even when","but not after"⟩
⟨EXCLUSIVE,"even when","except when"⟩
⟨EXCLUSIVE,"even when","except after"⟩
⟨EXCLUSIVE,"when","but not when"⟩
⟨EXCLUSIVE,"when","but not after"⟩
⟨EXCLUSIVE,"when","except when"⟩
⟨EXCLUSIVE,"when","except after"⟩
⟨EXCLUSIVE,"after","but not when"⟩
⟨EXCLUSIVE,"after","but not after"⟩
⟨EXCLUSIVE,"after","except when"⟩
⟨EXCLUSIVE,"after","except after"⟩

Figure A.8: The *only when* fragment

⟨EXCLUSIVE,“whether or not”,“only if”⟩
⟨EXCLUSIVE,“regardless of whether”,“only if”⟩
⟨EXCLUSIVE,“even if”,“only if”⟩
⟨SYNONYM,“whether or not”,“regardless of whether”⟩

Figure A.9: The *whether or not* fragment

⟨HYPERNYM,“before”,“even before”⟩
⟨EXCLUSIVE,“only before”,“even before”⟩
⟨EXCLUSIVE,“before”,“except before”⟩
⟨EXCLUSIVE,“only before”,“except before”⟩
⟨EXCLUSIVE,“even before”,“except before”⟩

Figure A.10: The *only before* fragment

⟨HYPERNYM,“because”,“for the reason that”⟩
⟨HYPERNYM,“because”,“on the grounds that”⟩
⟨HYPERNYM,“for the reason that”,“on the grounds that”⟩

Figure A.11: The *for the reason that* fragment

⟨HYPERNYM,“supposing that”,“if ever”⟩
⟨EXCLUSIVE,“presumably because”,“if”⟩
⟨EXCLUSIVE,“presumably because”,“in the hope that”⟩
⟨EXCLUSIVE,“if”,“in the hope that”⟩

Figure A.12: The *in the hope that* fragment

⟨HYPERNYM,“but”,“except”⟩

Figure A.13: The *except* fragment

Appendix B

Classification of discourse connectives using alternative similarity functions

This appendix provides details of experiments using two further distributional similarity functions to classify discourse connectives according to their semantic properties, as in Chapter 4. In the definitions of these functions below we assume that p and q are probability distributions corresponding to two connectives w_p and w_q, respectively.

The first function, $Jacc_t$, is a t-test weighted adaptation of the Jaccard coefficient (Curran and Moens, 2002a). In its basic form, the Jaccard coefficient is essentially a measure of how much two distributions overlap. The t-test variant weights co-occurrences by the strength of their collocation, using the following function:

$$wt(w_i,x) = \frac{P(w_i,x) - P(w_i)P(x)}{\sqrt{P(w_i)P(x)}} \tag{B.1}$$

(Here $P(x)$ is the probability of x, and $P(w_i,x)$ is the joint probability of w_i and x.) These weights are used to define the weighted version of the Jaccard coefficient, as shown in (B.2).

$$Jacc_t(p,q) = \frac{\sum_x min(wt(w_p,x), wt(w_q,x))}{\sum_x max(wt(w_p,x), wt(w_q,x))} \tag{B.2}$$

Because $Jacc_t$ has this inbuilt system for weighting co-occurrences on the strength of their collocation, it was not used in the experiments reported below which use just the most informative discourse markers co-occurrences. The second additional similarity function used was

the Euclidean distance function L_2, shown in (B.3), applied to probability distributions.

$$L_2(p,q) = \sqrt{\sum_x (p(x) - q(x))^2} \tag{B.3}$$

Results are reported using a co-occurrences variety of different word classes to construct the probability distributions. The following abbreviations are used for the different word classes:

VB:	non-auxiliary verbs
AUX:	auxiliary verbs
NN:	nouns (excluding pronouns)
PRP:	pronouns
JJ:	adjectives
RB:	adverbs
IN:	prepositions
DM:	discourse markers

The results obtained using *KL* and reported in Chapter 4 are also repeated for ease of comparison.

B.1 The polarity task

Distance function	Type of co-occurrences used as features								
	All POS	VB	AUX	NN	PRP	JJ	RB	IN	DM
KL	0.721	0.698	0.605	0.651	0.605	0.721	0.837	0.674	0.814
$Jacc_t$	0.744	0.744	0.651	0.721	0.605	0.721	0.767	0.721	0.814
L_2	0.744	0.698	0.512	0.628	0.651	0.651	0.767	0.721	0.744
Baseline	0.674								

Table B.1: Accuracy using the 1NN classifier on lexical co-occurrences

Classifier	Co-occurrences	Accuracy
Naive Bayes	All DMs	0.907
Naive Bayes	Most informative DMs	0.814
1NN with KL	Most informative DMs	0.698
1NN with L_2	Most informative DMs	0.721

Table B.2: Accuracy using Naive Bayes, and using the most informative discourse marker co-occurrences

B.2 The veridicality task

Distance function	Type of co-occurrences used as features								
	All POS	VB	AUX	NN	PRP	JJ	RB	IN	DM
KL	0.857	0.918	0.755	0.816	0.796	0.796	0.673	0.776	0.796
$Jacc_t$	0.755	0.878	0.673	0.857	0.735	0.755	0.714	0.776	0.837
L_2	0.816	0.816	0.673	0.837	0.735	0.816	0.693	0.816	0.837
Baseline	0.735								

Table B.3: Accuracy using the 1NN classifier on lexical co-occurrences

Classifier	Co-occurrences	Accuracy
Naive Bayes	All DMs	0.735
Naive Bayes	Most informative DMs	0.918
1NN with KL	Most informative DMs	0.776
1NN with L_2	Most informative DMs	0.857

Table B.4: Accuracy using Naive Bayes, and using the most informative discourse marker co-occurrences

B.3 The type task

Distance function	Type of co-occurrences used as features								
	All POS	VB	AUX	NN	PRP	JJ	RB	IN	DM
KL	0.645	0.677	0.548	0.677	0.613	0.613	0.742	0.710	0.801
$Jacc_t$	0.774	0.677	0.548	0.677	0.516	0.774	0.742	0.742	0.801
L_2	0.742	0.677	0.452	0.516	0.548	0.581	0.710	0.742	0.742
Baseline	0.581								

Table B.5: Accuracy using the 1NN classifier on lexical co-occurrences

Classifier	Co-occurrences	Accuracy
Naive Bayes	All DMs	0.581
Naive Bayes	Most informative DMs	0.935
1NN with *KL*	Most informative DMs	0.581
1NN with L_2	Most informative DMs	0.677

Table B.6: Accuracy using Naive Bayes, and using the most informative discourse marker co-occurrences

B.4 The direction task

Distance function	Type of co-occurrences used as features								
	All POS	VB	RB	AUX	NN	PRP	JJ	IN	DM
KL	0.811	0.676	0.784	0.811	0.730	0.865	0.757	0.757	0.784
$Jacc_t$	0.811	0.784	0.892	0.811	0.784	0.811	0.811	0.811	0.838
L_2	0.865	0.757	0.757	0.757	0.676	0.757	0.703	0.730	0.595
Baseline	0.811								

Table B.7: Accuracy using the 1NN classifier on lexical co-occurrences

Classifier	Co-occurrences	Accuracy
Naive Bayes	All DMs	0.838
Naive Bayes	Most informative DMs	0.973
1NN with KL	Most informative DMs	0.892
1NN with L_2	Most informative DMs	0.973

Table B.8: Accuracy using Naive Bayes, and using the most informative discourse marker co-occurrences

Appendix C

Eliciting judgements on the similarity of pairs of connectives

C.1 Instructions

The following instructions were presented to subjects before they were asked to judge the similarity of pairs of connectives.

Instructions

During the experiment you will see a number of words and phrases that can be used to connect sentences together. Examples of such words and phrases are shown below:

- Jim had a lot of money on him that day, **so** he went shopping.
- The software can generate realistic background images **so that** users can pretend they are somewhere else.
- Bob shouted very loudly, **but** nobody heard him.
- It's a fairly good piece of work, **considering that** you have been under a lot of pressure lately.
- I wouldn't vote for Smith **even if** you gave me a thousand pounds.

In the experiment you will be presented with pairs of such connective words and phrases. Your task is to judge how similar the two connectives are in meaning. For example, you may be asked to judge the similarity of the following two connectives:

- *Something happened* **when** *something else happened*
- *Something happened* **while** *something else happened*

You will do this by entering a number between **0** and **5**, with **0** indicating the connectives are **not similar** in meaning at all, and **5** indicating the connectives are **very similar** in meaning. So for the pair *when* and *while* you might assign a score towards the upper end of the scale.

On the other hand, you might instead see 'because' and 'whereas':

- *Something happened* **because** *something else happened*
- *Something happened* **whereas** *something else happened*

In this case you might assign a low score as the two words do not seem very similar.

As one last example, suppose you are presented with:

- *Something happened* **after** *something else happened*
- *Something happened* **before** *something else happened*

In this case **after** and **before** do seem somewhat similar, as they both indicate the temporal ordering between events. However because they indicate opposite temporal orderings, you might only assign them an average score of about **2** or **3**.

A word of caution: it is the similarity in **meaning** of the connectives that we are interested in. For example, *so* and *so that* don't mean the same thing, despite the similarity in their spellings!

Your personal details

Before the actual experiment begins, you'll see a form asking for details about yourself (this is the first thing you will see once you've pressed the start button below). I'd be grateful if you'd give a valid email address so that we can contact you if we have any questions about your answers, and so that we can mail you with information about the purpose of the experiment once it is completed. Also don't forget that we need valid email address for entering you in the prize draw.

Please be careful to fill in the Personal Details questionnaire correctly, as otherwise we will have to discard your responses. We ask you to supply the following information:

- your name and email address;

- your age and sex;
- whether you are right or left handed (based on the hand you prefer to use for writing);
- the academic subject you study or have studied (or your current occupation in case you haven't attended university);
- In the field marked 'Region' I'd like you to give me an indication of the region you grew up in, so that we have an idea of the type of English you speak (I'd like this information in case there are differences between dialects!).

The personal data you give me is used only for scientific purposes. I will not give any of this information to anyone else, and nor will I report any information in any way that can be identified with you.

And finally...

Taking part in this experiment is entirely voluntary! Obviously I'd be grateful if you stayed the course, but of course you are at liberty to break off at any point during the experiment. If you choose to do so, please skip to the bottom of the page and press the 'submit' button so that your answers will be recorded.

Once again, thanks for your interest in taking part, and have fun! You can start the experiment proper by pressing on the 'Start' button below. The page may be slow to load. If so, your patience is appreciated.

C.2 Results

SYNONYMOUS connectives	Mean	Std Dev	High	Low
although–despite the fact that	4.125	1.114	5	1
now–now that	3.128	1.750	5	0
but–yet	3.897	1.142	5	1
considering that–given that	3.875	1.223	5	0
or else–or	3.154	1.829	5	0
despite the fact that–even though	4.675	0.616	5	3
on the assumption that–assuming that	4.462	0.884	5	2

considering that–seeing as	4.300	1.042	5	1
regardless of whether–whether or not	4.821	0.506	5	3
just as–the way	3.051	1.669	5	0
although–even though	4.150	0.893	5	2
seeing as–given that	3.945	1.191	5	1
HYPONYMOUS connectives	Mean	Std Dev	High	Low
notwithstanding that–even though	3.900	1.317	5	0
if–on condition that	4.462	1.022	5	0
if–if only	2.925	1.474	5	0
lest–in case	3.925	1.366	5	0
as soon as–the moment	4.564	0.754	5	2
if–if ever	3.180	1.144	5	1
and–whereas	1.513	1.295	5	0
supposing that–if ever	2.850	1.369	5	0
although–notwithstanding that	3.375	1.372	5	0
for–because	3.256	1.743	5	0
if–on the assumption that	3.600	1.336	5	0
if–assuming that	3.641	1.287	5	0
CONTINGENTLY SUBSTITUTABLE connectives	Mean	Std Dev	High	Low
much as†–yet	0.795	1.005	4	0
but then–much as†	0.744	0.850	3	0
but–despite the fact that	1.800	1.488	5	0
but not when–by the time	1.154	1.065	3	0
in that–seeing as	2.975	1.441	5	0
given that–in that	3.103	1.429	5	0
but not when–except since	2.180	1.335	4	0
if–only if	3.154	1.288	5	0
for fear that–regardless of whether	0.769	0.902	3	0
as–in that	2.205	1.341	4	0
and–or	0.615	1.016	5	0

for–insofar as	1.923	1.244	4	0
EXCLUSIVE connectives	Mean	Std Dev	High	Low
but–only if	0.775	0.800	3	0
for fear that–seeing as	0.949	0.999	4	0
but–now that	0.949	0.972	4	0
just as–supposing that	0.850	1.099	4	0
for fear that–until	0.564	0.821	3	0
although–except when	1.154	1.368	5	0
the way–as	2.308	1.704	5	0
and–assuming that	0.615	0.748	2	0
only after–whether or not	0.550	0.986	5	0
just as–now that	1.667	1.305	5	0
considering that–in order that	1.333	1.344	5	0
only when–so that	1.308	1.196	4	0

† *much as* cannot easily connect events, which may have caused subjects difficulties in rating these items.

Appendix D

A hierarchical clustering of discourse connectives

In Chapter 5 we found that distributional similarity correlated positively with semantic similarity. This appendix provides a hierarchy of discourse connectives which was constructed on the basis of distributional similarity. We provide the hierarchy purely because it may be useful as a descriptive resource. That is, we make no claims or hypotheses about the clusters it contains, other than that they summarise the distributional similarities between a large number of discourse connectives.

This hierarchy was produced automatically using agglomerative hierarchical clustering (Jain et al., 1999). To do this, a symmetric distance function was defined by applying the Kullback-Leibler divergence function, and taking the average of applying with arguments in both possible orders:

$$distance(p,q) = \frac{D(p||q) + D(q||p)}{2} \tag{D.1}$$

Using Lee's (1999) skewed version of KL divergence ensured this was always defined (another solution would be to use a clustering method which does not calculate distances between individual word distributions (Pereira et al., 1993)). The distances between two clusters was taken to be the average of the distances between their members, although alternative methods are possible (see Schulte im Walde, 2003).

The size of the hierarchy necessitates splitting it into several diagrams, and it is presented in Figures D.1, D.2, D.3, D.4 and D.5. To obtain the complete hierarchy, root nodes with labels of the form **CLUSTER**$**$ should be adjoined into the corresponding leaf nodes. The labels of subclusters of the hierarchy contain numbers which indicate the order in which the subclusters

were created. As a result, a lower cluster number indicates a smaller distance between its constituents. Note that the distributional similarity scores are based on co-occurrences with discourse markers. Quite different hierarchies might result if different types of co-occurrences (e.g. co-occurrences with verbs) are used instead.

It can be observed that the TOP node of the hierarchy has two daughter nodes, one of which contains just the two discourse connectives *which was why* and *which is why*. This indicates that these connectives have quite different distributions from all other ones. Similarly, other connectives close to the top node also have quite distinct distributions, e.g. *except before, except since, but not when, by the time* and *but not after*. At the other extreme, cluster C5 (in Figure D.5) contains four connectives with very similar co-occurrence distributions: *so, because, as* and *for*. The fact that this cluster was created fifth (which is deducible from its label "C5") indicates that these are among the most distributionally similar connectives in the hierarchy.

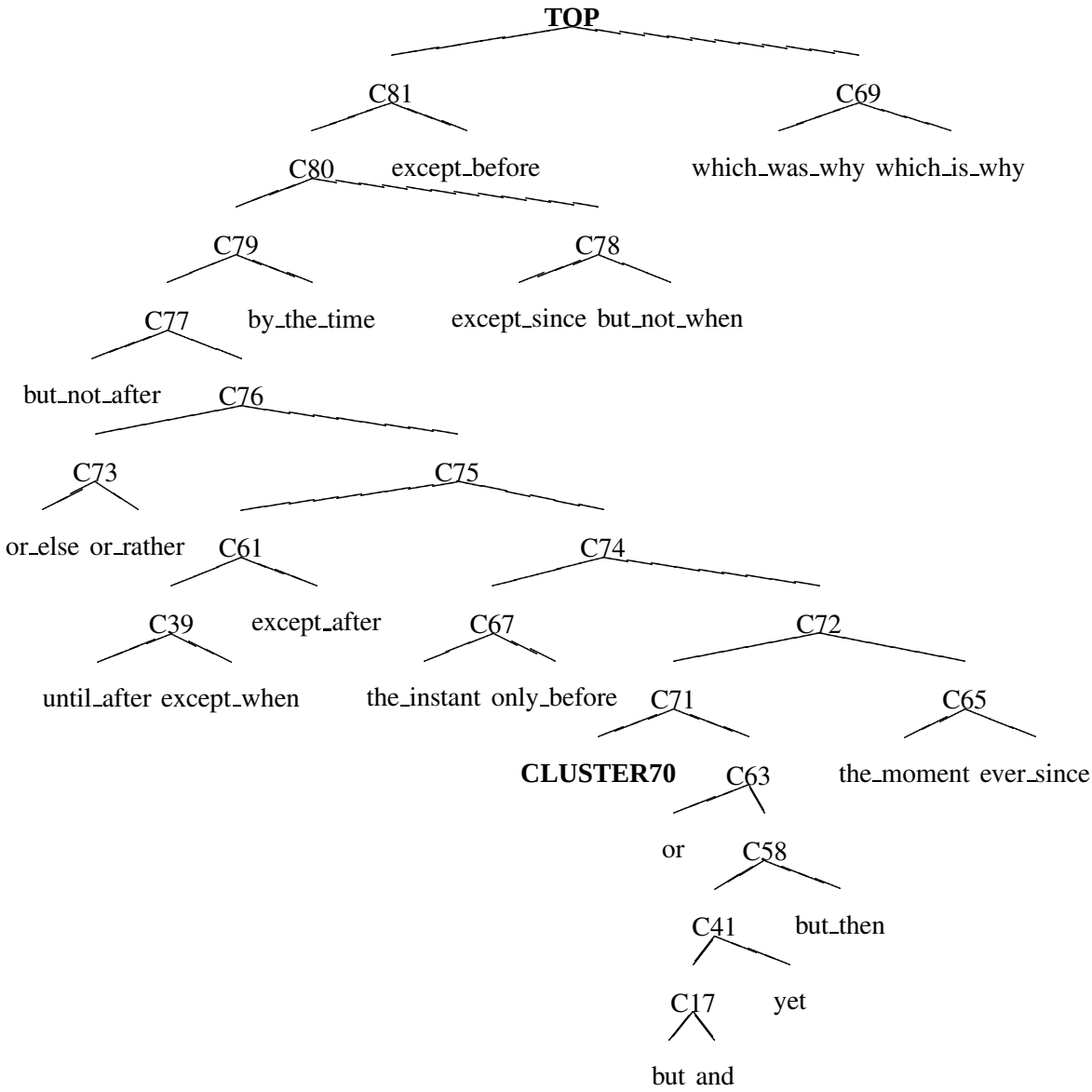


Figure D.1: Top levels of the hierarchy

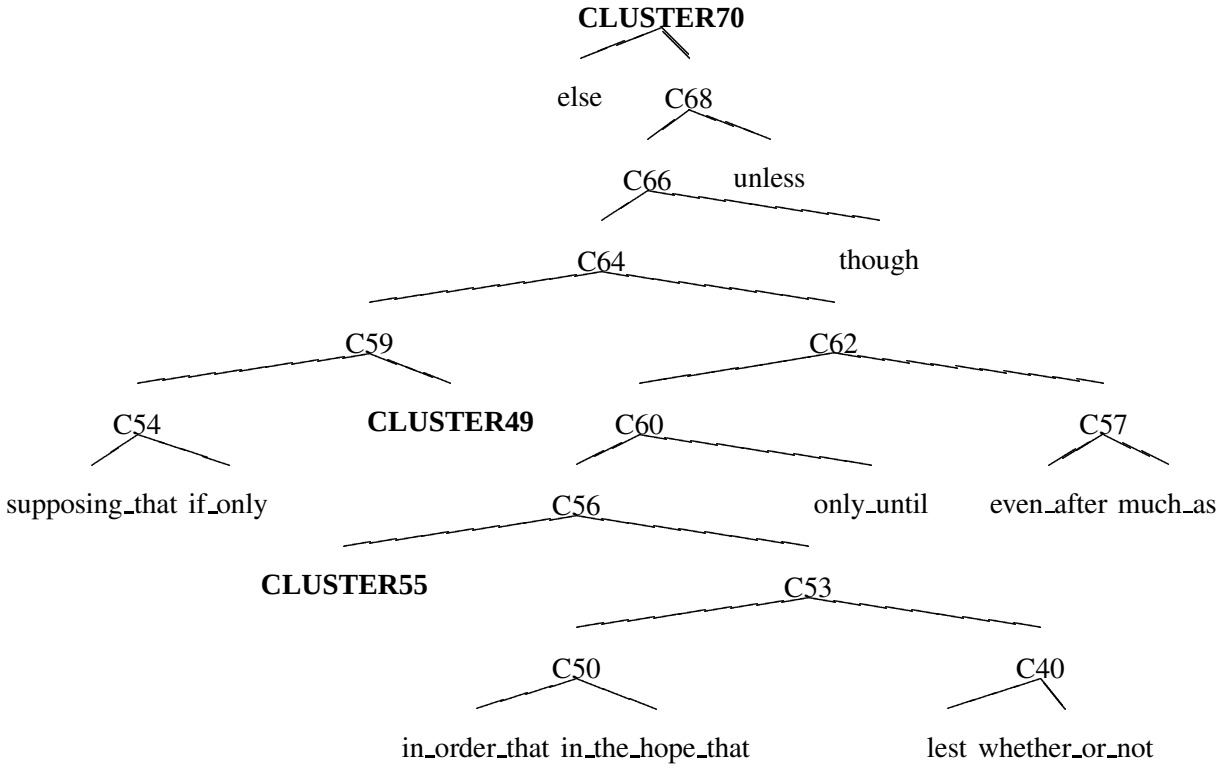


Figure D.2: Subcluster 70 of the hierarchy

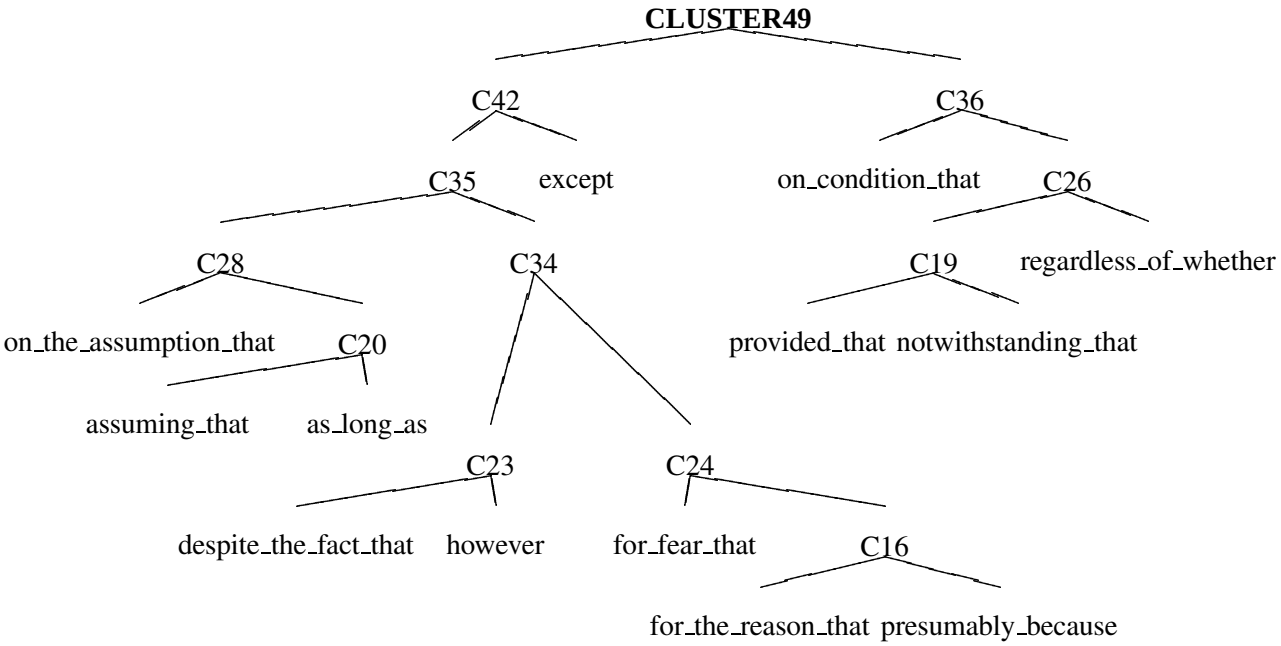


Figure D.3: Subcluster 49 of the hierarchy

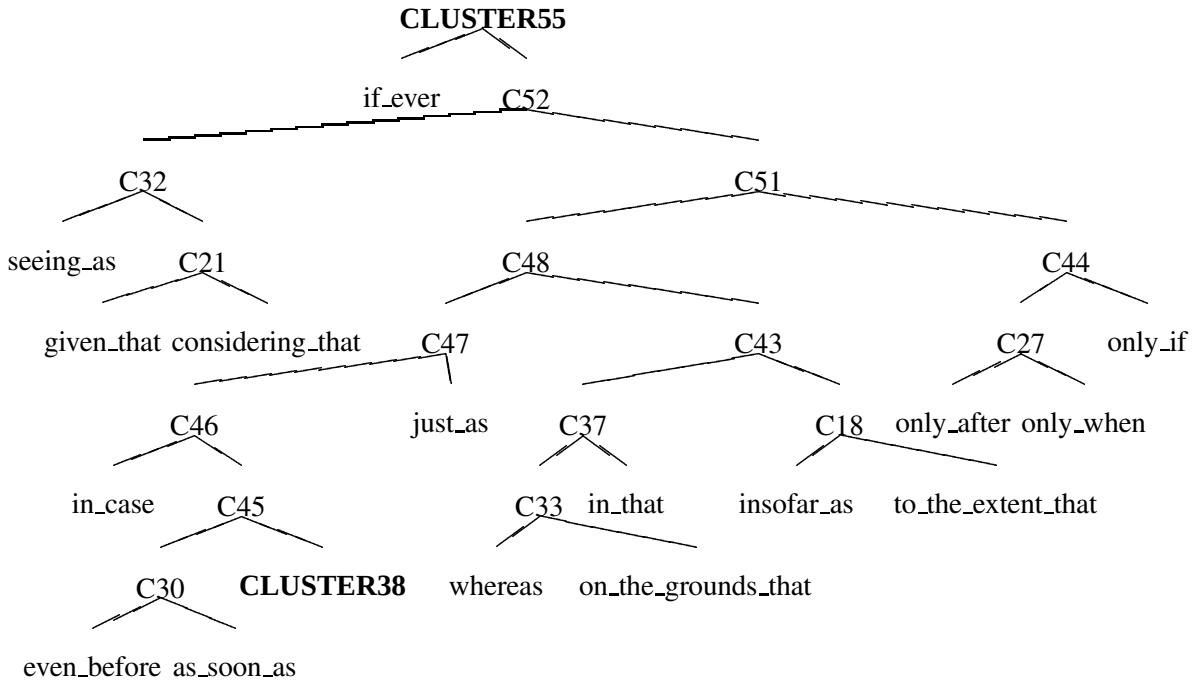


Figure D.4: Subcluster 55 of the hierarchy

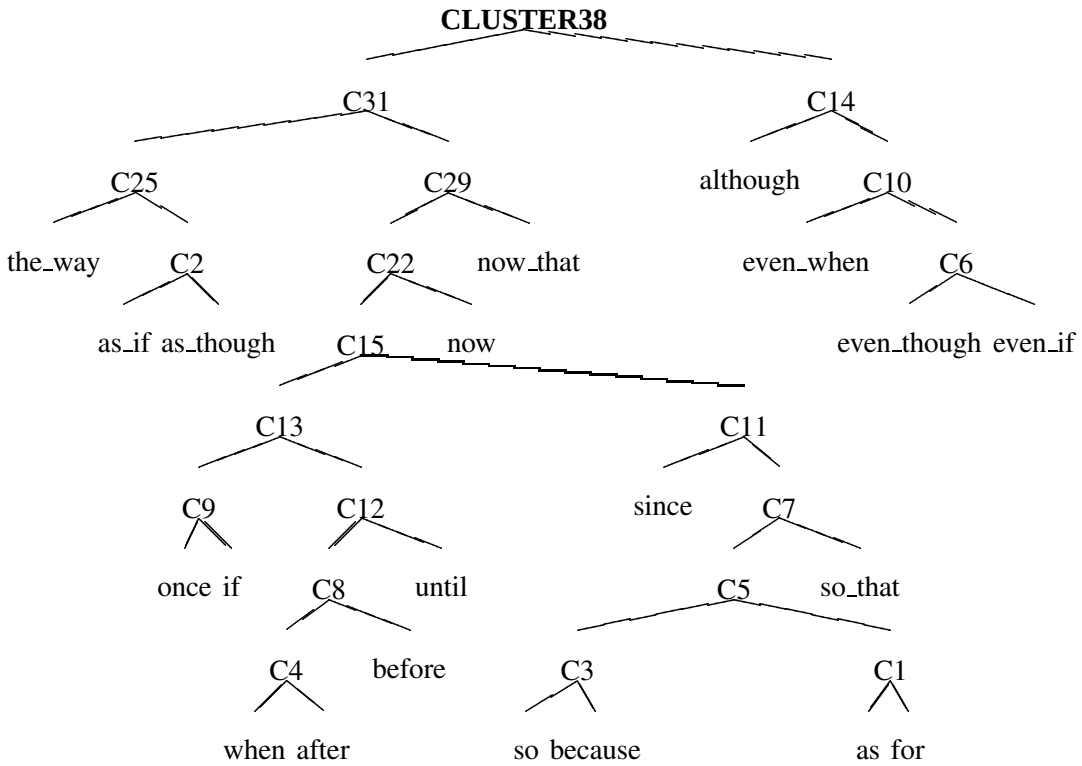


Figure D.5: Subcluster 38 of the hierarchy

Appendix E

Ensembles for predicting substitutability

The ensembles of classifiers used in Chapter 6 were each constructed with a particular evaluation metric in mind. Given a metric, the validation data was used to select the best 20 individual classifiers, with the constraint that the beam size was fixed at 5. Given these 20-best lists, the top N classifiers were used to construct an ensemble ($1 \leq N \leq 20$). In this Appendix we list the top 20 performing classifiers for each metric. Each classifier is specified as a list of triples $\langle C, F, S \rangle$, where C is a type of lexical co-occurrence, F is a type of distributional function (either $D(X||Y)$ or $V(X,Y)$), and S specifies the symmetric meta-function that was applied to the distributional function, which could be one of:

- "Average", e.g. $\frac{D(X||Y)+D(Y||X)}{2}$.
- "Average of logarithms", e.g. $\frac{\log(D(X||Y))+\log(D(Y||X))}{2}$. This symmetrisation is more appropriate than the previous one for distributional functions with a right skewed distribution.
- "Difference squared", e.g. $(V(X,Y) - V(Y,X))^2$. This symmetrisation aims to exploit the discovery in Chapter 5 that the asymmetry of $V(p,q)$ can be used to predict substitutability.
- "Difference squared over average", e.g. $(V(X,Y) - V(Y,X))^2 / \frac{V(X,Y)+V(Y,X)}{2}$. Similar to the previous one, except normalising for magnitude.

When a classifier is specified using only one triple, we have what we describe in Chapter 6 as a "simple data representation". When there is more than one triple, we have a "compound data representation".

1	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
2	⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
3	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
4	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩
5	⟨verbs,$D(X\|\|Y)$,average of logs⟩
6	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
7	⟨all words,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
8	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩
9	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$V(X,Y)$,average⟩, ⟨all words,$V(X,Y)$,average⟩
10	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨discourse markers,$D(X\|\|Y)$,difference squared over average⟩
11	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
12	⟨discourse markers,$V(X,Y)$,average⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
13	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨discourse markers,$V(X,Y)$,average⟩
14	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
15	⟨all words,$D(X\|\|Y)$,average of logs⟩
16	⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
17	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
18	⟨discourse markers,$V(X,Y)$,difference squared over average⟩
19	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩
20	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$V(X,Y)$,average⟩

Figure E.1: Best performing classifiers on the accuracy metric

1	⟨ all words,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
2	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
3	⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
4	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
5	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨discourse markers,$D(X\|\|Y)$,difference squared over average⟩
6	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
7	⟨all words,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩, ⟨all words,$V(X,Y)$,difference squared over average⟩
8	⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
9	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
10	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
11	⟨all words,$D(X\|\|Y)$,average of logs⟩
12	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$V(X,Y)$,average⟩, ⟨all words,$V(X,Y)$,average⟩
13	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩
14	⟨discourse markers,$D(X\|\|Y)$,average⟩
15	⟨verbs,$D(X\|\|Y)$,average of logs⟩
16	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨discourse markers,$V(X,Y)$,average⟩, ⟨discourse markers,$V(X,Y)$,difference squared over average⟩
17	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
18	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩
19	⟨discourse markers,$V(X,Y)$,average⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
20	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$V(X,Y)$,average⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩

Figure E.2: Best performing classifiers on the Obtained Information metric

1	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
2	⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
3	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
4	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩
5	⟨verbs,$D(X\|\|Y)$,average of logs⟩
6	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
7	⟨all words,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
8	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨discourse markers,$D(X\|\|Y)$,difference squared over average⟩
9	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$V(X,Y)$,average⟩, ⟨all words,$V(X,Y)$,average⟩
10	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩
11	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
12	⟨discourse markers,$V(X,Y)$,average⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
13	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
14	⟨all words,$D(X\|\|Y)$,average of logs⟩
15	⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
16	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
17	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
18	⟨discourse markers,$V(X,Y)$,difference squared over average⟩
19	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩
20	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$V(X,Y)$,average⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩

Figure E.3: Best performing classifiers on the Relative Information metric

1	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
2	⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
3	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
4	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨discourse markers,$D(X\|\|Y)$,difference squared over average⟩
5	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
6	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩
7	⟨all words,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩
8	⟨verbs,$D(X\|\|Y)$,average of logs⟩
9	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$V(X,Y)$,average⟩, ⟨all words,$V(X,Y)$,average⟩
10	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩
11	⟨discourse markers,$V(X,Y)$,average⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
12	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
13	⟨all words,$D(X\|\|Y)$,average of logs⟩
14	⟨verbs,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
15	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
16	⟨discourse markers,$V(X,Y)$,average⟩, ⟨all words,$D(X\|\|Y)$,average of logs⟩
17	⟨discourse markers,$D(X\|\|Y)$,average of logs⟩, ⟨discourse markers,$V(X,Y)$,average⟩
18	⟨discourse markers,$V(X,Y)$,difference squared over average⟩
19	⟨discourse markers,$D(X\|\|Y)$,average⟩
20	⟨all words,$D(X\|\|Y)$,average of logs⟩, ⟨all words,$V(X,Y)$,average⟩, ⟨all words,$V(X,Y)$,difference squared over average⟩

Figure E.4: Best performing classifiers on the kappa metric

Bibliography

Aha, D. W., Kibler, D. F., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Alfonseca, E. and Manandhar, S. (2002). Improving an ontology refinement method with hyponymy patterns. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain.

Allen, J. (2003). Post-editing. In Somers, H., editor, *Computers and Translation: A Translators Guide*, volume 35 of *Benjamins Translation Library*. John Benjamins, Amsterdam.

Althaus, E., Karamanis, N., and Koller, A. (2004). Computing locally coherent discourses. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht.

Asher, N. and Lascarides, A. (1995). Lexical disambiguation in a discourse context. *Journal of Semantics*, 12(1):69–108.

Asher, N. and Lascarides, A. (1998). Questions in dialogue. *Linguistics and Philosophy*, 23(2):237–309.

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge.

Baldridge, J. (2002). *Lexically Specified Derivational Control in Combinatory Categorial Grammar*. PhD thesis, University of Edinburgh.

Baldwin, T. and Bond, F. (2003). Learning the countability of English nouns from corpus data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 463–470, Sapporo, Japan.

Ballard, D., Conrad, R., and Longacre, R. (1971). The deep and surface grammar of interclausal relations. *Foundations of Language*, 4:70–118.

Bangalore, S. and Rambow, O. (2000). Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany.

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Barzilay, R. and Lee, L. (2004). Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2004)*.

Basili, R., Vindigni, M., and Zanzotto, F. M. (2004). Integrating semantic lexicons and domain ontologies. In *Proceedings of Workshop OntoLex 2004*, Lisbon, Portugal.

Berger, D., Reitter, D., and Stede, M. (2002). XML/XSL in the dictionary: the case of discourse markers. In *Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002)*, Taipei.

Bestgen, Y., Degand, L., and Spooren, W. (2003). On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: An exploratory study. In *Proceedings of the MAD'03 workshop on Multidisciplinary Approaches to Discourse*.

Blakemore, D. and Carston, R. (2005). The pragmatics of sentential coordination with 'and'. *Lingua, special issue on Coordination: Syntax, Semantics and Pragmatics*, 115(4):569–589.

Boguraev, B. and Pustejovsky, J., editors (1996). *Corpus Processing for Lexical Acquisition*. MIT Press, Cambridge, MA, USA.

Brew, C. and Schulte im Walde, S. (2002). Spectral clustering for german verbs. In Hajic and Matsumoto, editors, *Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing*, pages 117–124, Philadelphia, USA.

Briscoe, T., de Paiva, V., and Copestake, A. (1993). *Inheritance, defaults, and the lexicon*. Cambridge University Press, Cambridge.

Budanitsky, A. and Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh.

Burnard, L., editor (1995). *Users Reference Guide for the British National Corpus*. Oxford University Computing Service.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics*, Maryland, USA.

Caraballo, S. A. and Charniak, E. (1999). Determining the specificity of nouns from text. In *Proceedings of the joint SIGDAT conference on Empirical Methods in NLP and Very Large Corpora*, Maryland, USA.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Caron, J., Micko, H. C., and Thüring, M. (1988). Conjunctions and recall of composite sentences. *Journal of Memory and Language*, 27:309–323.

Cederberg, S. and Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*, pages 111–118, Edmonton, Canada.

Charles, W. G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics*, 21:505–524.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, Seattle, Washington, USA.

Cimiano, P., Hotho, A., and Staab, S. (2004). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In *Proceedings of the 16th European Conference on Artificial Intelligence*, Valencia, Spain.

Cohen, R. (1984). A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 251–258, Stanford, CA, USA.

Corston-Oliver, S. H. (1998). *Computing Representations of the Structure of Written Discourse*. PhD thesis, University of California Santa Barbara.

Creswell, C. (2003). *Syntactic form and discourse function in natural language generation*. PhD thesis, Linguistics, University of Pennsylvania.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

Crystal, D. (1997). *A dictionary of linguistics and phonetics (4th edition)*. Blackwell Publishers, Oxford, UK.

Cucchiarelli, A., Luzi, D., and Velardi, P. (1998). Automatic semantic tagging of unknown proper names. In Boitet, C. and Whitelock, P., editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 286–292, San Francisco, California. Morgan Kaufmann Publishers.

Curran, J. (2002). Ensemble methods for automatic thesaurus extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 222–229, Philadelphia, PA, USA.

Curran, J. R. and Moens, M. (2002a). Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, pages 59–67, Philadelphia, PA, USA.

Curran, J. R. and Moens, M. (2002b). Scaling context space. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Philadelphia, PA, USA.

Curran, J. R. and Osborne, M. (2002). A very very large corpus does not always yield reliable estimates. In *Proceedings of the 6th Conference on Natural Language Learning*, Taipei, Taiwan.

Daelemans, W., De Smedt, K., and Gazdar, G. (1992). Inheritance in natural language processing. *Computational Linguistics*, 18(2):205–218.

Dagan, I., Lee, L., and Pereira, F. C. C. (1999). Similarity-based models of word co-occurrence probabilities. *Machine Learning*, 34:43–69.

Delin, J., Scott, D., and Hartley, A. (1996). Pragmatic congruence through language-specific mappings from semantics to syntax. In *Proceedings of the 16th Conference on Computational Linguistics*, Copenhagen, Denmark.

Di Eugenio, B., Moore, J. D., and Paolucci, M. (1997). Learning features that predict cue usage. In *Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL97)*, Madrid, Spain.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857.

Durieux, G., Daelemans, W., and Gillis, S. (1999). On the arbitrariness of lexical categories. In Van Eynde, F., Schuurman, I., and Schelkens, N., editors, *Computational Linguistics in The Netherlands 1998*, pages 19–36, Amsterdam. Rodopi.

Elhadad, M. and McKeown, K. (1990). Generating connectives. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 97–101, Helsinki, Finland.

Eyheramendy, S., Lewis, D. D., and Madigan, D. (2003). On the naive bayes model for text categorization. In *Proceedings of the Ninth International Workshop on AI and Statistics*, Florida, US.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford.

Forbes, K., Miltsakaki, E., Prasad, R., Sarkar, A., Joshi, A., and Webber, B. (2003). D-LTAG system — discourse parsing with a lexicalized tree-adjoining grammar. *Journal of Language, Logic and Information*, 12(3).

Forbes, K. and Webber, B. (2002). A semantic account of adverbials as discourse connectives. In *Proceedings of the ACL workshop on Discourse and Dialogue (SIGDIAL)*.

Forbes, K. M. (2003). *Discourse Semantics of S-Modifying Adverbials*. PhD thesis, University of Pennsylvania.

Ganter, B. and Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin.

Girju, R. and Woods, K. (2005). Exploring contingency discourse relations. In Bunt, H., editor, *Proceedings of the 6th International Workshop on Computational Semantics (IWCS-6)*, The Netherlands.

Glasbey, S. (1995). 'when', discourse relations and the thematic structure of events. In *Proceedings on the Conference on Time, Space, and Movement*, University of Toulouse.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Speech Acts*, pages 41–58. Academic Press, New York.

Grimes, J. (1975). *The Thread of Discourse*. Morton, The Hague.

Grosz, B. J. and Sidner, C. J. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–203.

Grote, B., Lenke, N., and Stede, M. (1997). Mar(k)ing concessions in English and German. *Discourse Processes*, 24(1):87–118.

Grote, B. and Stede, M. (1998). Discourse marker choice in sentence planning. In Hovy, E., editor, *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 128–137, New Brunswick, New Jersey. Association for Computational Linguistics.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman.

Hamann, C. (1989). English temporal clauses in a reference frame model. In Schopf, A., editor, *Essays on Tensing in English, Vol II: Time, Text and modality*, pages 31–153. Max Niemeyer, Tübingen.

Harris, Z. S. (1970). *Papers in structural and transformational linguistics*. Reidel, Dordrecht.

Hearst, M. (1998). Automated discovery of WordNet relations. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*. MIT Press.

Hearst, M. and Schütze, H. (1996). Customizing a lexicon to better suit a computational task. In Boguraev, B. and Pustejovsky, J., editors, *Corpus Processing for Lexical Acquisition*. MIT Press.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference in Computational Linguistics*, Nantes, France.

Heinamaki, O. (1972). 'before'. In *Papers from the Eighth Regional Meeting of the Chicago Linguistic Society*, pages 139–151. Chicago Linguistic Society.

Henderson, J. C. and Brill, E. (1999). Exploiting diversity in natural language processing: Combining parsers. In *Proceedings of the Fourth Conference on Empirical Methods in Natural Language Processing*, College Park, Maryland.

Hindle, D. (1990). Noun classification from predicate argument structures. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, Pennsylvania, USA.

Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Hirst, G. and St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In (Fellbaum, 1998), pages 305–332.

Hobbs, J., Stickel, M., Appelt, D., and Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142.

Hobbs, J. A. (1985). On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University.

Hobbs, J. R. (1990). *Literature and Cognition*. Lecture Notes 2. CSLI Publications.

Hovy, E. H., Lavid, J., Maier, E., Mittal, V., and Paris, C. L. (1992). Employing knowledge resources in a new text planning architecture. In Dale, R., Hovy, E., Rösner, D., and Stock, O., editors, *Aspects of Automated Natural Language Generation*, Springer Verlag Lecture Notes in AI number 587, pages 57–72, Heidelberg, Germany.

Howell, D. C. (2002). *Statistical Methods for Psychology*. Duxbury, California, USA, fifth edition.

Huddleston, R. and Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge University Press, Cambridge, UK.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Oxford University Press, Oxford. 1999 edition.

Hutchinson, B. (2004a). Acquiring the meaning of discourse markers. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 685–692, Barcelona, Spain.

Hutchinson, B. (2004b). Mining the web for discourse markers. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 407–410, Lisbon, Portugal.

Hutchinson, B. (2004c). What we can learn from co-occurrences of discourse connectives: A corpus-based study. Talk given at the CSDL 2004 Conference on Conceptual Structure, Discourse, & Language. Edmonton, Canada.

Hutchinson, B. (2005a). MDL-based acquisition of substitutability relationships between discourse connectives. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*, pages 1690–1691, Edinburgh, UK.

Hutchinson, B. (2005b). Modelling the similarity of discourse connectives. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society (CogSci2005)*, pages 1012–1017, Stresa, Italy.

Hutchinson, B. (2005c). Modelling the substitutability of discourse connectives. In *Proceedings of the 43rd Annual Meeting of The Association for Computational Linguistics*, pages 149–156, Ann Arbor, USA.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Survey*, 31(3):264–323.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.

Joshi, A. (1987). An introduction to tree adjoining grammar. In Manaster-Ramer, A., editor, *Mathematics of Language*, pages 87–114. John Benjamins, Amsterdam.

Joshi, A. K. and Vijay-Shanker, K. (2001). Compositional semantics for lexicalized tree-adjoining grammars. In Bunt, H., Muskens, R., and Thijsse, E., editors, *Computing Meaning, Studies in Linguistics and Philosophy*, volume 77. Kluwer Academic Press, Dodrecht.

Kallmeyer, L. and Joshi, A. (2003). Factoring predicate argument and scope semantics: Underspecified semantics with LTAG. *Research on Language and Computation*, 1(1–2):3–58.

Karttunen, L. (1973). Presuppositions of compound sentences. *Linguistic Inquiry*, 4:169–193.

Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar*. CSLI publications, Stanford.

Kehler, A., Appelt, D., Taylor, L., and Simma, A. (2004). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2004)*, Boston, MA.

Keller, F. (2000). *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. PhD thesis, University of Edinburgh.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Kilgarriff, A. and Grefenstette, G. (2003a). *Computational Linguistics: Special Issue on the Web as Corpus*, volume 29:3. MIT Press.

Kilgarriff, A. and Grefenstette, G. (2003b). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–348.

Knight, K. and Chander, I. (1994). Automated postediting of documents. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Knott, A. (1996). *A data-driven methodology for motivating a set of coherence relations*. PhD thesis, University of Edinburgh.

Knott, A. (2001). Semantic and pragmatic relations and their intended effects. In Sanders, T., Schilperoord, J., and Spooren, W., editors, *Text representation: linguistic and psycholinguistic aspects*, pages 127–151. Benjamins, Amsterdam.

Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Knott, A. and Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30:135–175.

Kononenko, I. and Bratko, I. (1991). Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6:67–80.

Krippendorf, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverley Hills, CA, USA.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Langkilde, I. and Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*.

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Meeting of the Association of Computational Linguistics*, pages 545–552, Sapporo, Japan.

Lapata, M. and Lascarides, A. (2002). A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):263–317.

Lapata, M. and Lascarides, A. (2004). Inferring sentence-internal temporal relations. In *In Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting*, Boston, MA.

Lascarides, A. and Asher, N. (1993). Temporal interpretation, discourse relations and common sense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Lascarides, A. and Asher, N. (1999). Cognitive states, discourse stucture, and the content of dialogue. In *Proceedings to Amsterlogue*.

Lascarides, A. and Asher, N. (2004). Imperatives in dialogue. In Kuehnlein, P., Rieser, H., and Zeevat, H., editors, *The Semantics and Pragmatics of Dialogue for the New Millenium*. Benjamins.

Lascarides, A. and Oberlander, J. (1992). Abducing temporal discourse. In Dale, R., Hovy, E., Rosner, D., and Stock, O., editors, *Aspects of Automated Natural Language Generation*, pages 167–182. Springer Verlag.

Lascarides, A. and Oberlander, J. (1993). Temporal coherence and defeasible knowledge. *Theoretical Linguistics*, 19(1):1–35.

Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In (Fellbaum, 1998), pages 265–283.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings the 27th Annual Meeting of the Association for Computational Linguisticsof (ACL 1999)*, Maryland, USA.

Lee, L. (2001). On the effectiveness of the skew divergence for statistical language analysis. *Artificial Intelligence and Statistics*, pages 65–72.

Li, H. (2002). Word clustering and disambiguation based on co-occurrence data. *Natural Language Engineering*, 8(1):25–42.

Li, H. and Abe, N. (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pages 768–773.

Lin, D. (1998b). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Litman, D. J. (1996). Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53–94.

Longacre, R. E. (1983). *The grammar of discourse*. Plenum, New York.

Louwerse, M. M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, 12(3):291–315.

Lynn, C. and Marcu, D. (2001). Discourse tagging manual. Technical Report ISI-TR-545, ISI.

Mani, I. and Wilson, G. (2000). Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 69–76, New Brunswick, New Jersey.

Mann, W. and Thompson, S. A. (1987). Rhetorical structure theory: Towards a functional theory of text organization. Technical Report ISI/RS-87-190, USC Information Sciences Institute.

Mann, W. C., Matthiessen, C. M. I. M., and Thompson, S. A. (1992). Rhetorical structure theory and text analysis. In Mann, W. C. and Thompson, S. A., editors, *Discourse Description: Diverse linguistic analyses of a fund-raising text*, pages 39–78. John Benjamins, Amsterdam.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT, Cambridge, Massachusetts.

Marcu, D. (1997). The rhetorical parsing of natural language texts. In *The Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97/EACL'97)*, pages 96–103, Madrid, Spain.

Marcu, D. (1998). A surface based approach to identifying discourse markers and elementary textual units in unrestricted texts. In *Proceedings of the COLING/ACL workshop on Discourse Relations and Discourse Markers*.

Marcu, D. (1999). A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 365–372, Maryland, USA.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, Cambridge, MA, USA.

Marcu, D. and Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19.

Martin, J. (1992). *English Text: System and Structure*. Benjamin, Amsterdam.

McDonald, S. (2000). *Environmental determinants of lexical processing effort*. PhD thesis, University of Edinburgh.

McKeown, K. R. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, Cambridge, England.

Merlo, P. and Stevenson, S. (2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Miller, G. A. (1990). Wordnet: An online lexical database. *International Journal of Lexicography*, 3(4).

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The Penn Discourse Treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon.

Moens, M. and Steedman, M. (1988). Temporal ontology and temporal reference. *Computational Linguistics: Special issue on tense and aspect*, 14:15–28.

Moore, J. D. and Pollack, M. E. (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Moser, M. and Moore, J. (1995). Using discourse analysis and automatic text generation to study discourse cue usage. In *Proceedings of the AAAI 1995 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 92–98.

Noordman, L. and Vonk, W. (1997). The different functions of a conjunction in constructing a representation of the discourse. In Costermans, J. and Fayol, M., editors, *Processing Interclausal relationships: Studies in the production and comprehension of texts*, pages 75–93. Lawrence Erlbaum, Mahawah, NJ.

Oates, S. L. (2000). Multiple discourse marker occurrence: Creating hierarchies for natural language generation. In *Proceedings of 6th Applied NLP Conference*.

Oates, S. L. (2001). Generating multiple discourse markers in text. Master's thesis, ITRI, University of Brighton.

Oberlander, J. and Knott, A. (1995). Issues in cue phrase implicature. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*.

Oberlander, J. and Lascarides, A. (1991). Discourse generation, temporal constraints and defeasible reasoning. In *Proceedings of the AAAI Fall Symposium on Discourse Structure in Interpretation and Generation*, Asilomar, California.

Osborne, M. and Baldridge, J. (2004). Ensemble-based active learning for parse selection. In *Proceedings of NAACL-04*, Boston, USA.

Padó, S. and Lapata, M. (2003). Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 126–135, Sapporo, Japan.

Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 183–190, Ohio, USA.

Petersen, W. (2001). A set-theoretical approach for the induction of inheritance hierarchies. *Language and Computation*, 1(3):1–14.

Pollard, C. J. and Sag, I. A. (1987). *Information-based Syntax and Semantics*. CSLI Publications, Stanford University. Vol. 1. CSLI Lecture Notes, number 13.

Pollard, C. J. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson, London.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Power, R., Scott, D., and Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics*, 29(4):211–260.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge/London.

Quinlan, J. R. and Rivest, R. L. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, 80:227–248.

Resnik, P. (1995). Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

Resnik, P. and Diab, M. (2000). Measuring verb similarity. In *Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society (COGSCI2000)*, Philadelphia, US.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica*, 14:465–471.

Roberts, C. (1989). Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy*, 12:683–721.

Rösner, D. and Stede, M. (1992). Customizing RST for the automatic production of technical manuals. In Dale, R., Hovy, E., Rösner, D., and Stock, O., editors, *Aspects of automated natural language generation. Proceedings of the 6th international workshop on natural language generation*, Berlin/Heidelberg. Springer.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Computational Linguistics*, 8:627–633.

Rydin, S. (2002). Building a hyponymy lexicon with hierarchical structure. In *Proceedings of the SIGLEX Workshop on Unsupervised Lexical Acquisition*, pages 26–33.

Sacco, K., Bucciarelli, M., and Adenzato, M. (2001). Mental models and the meaning of connectives: a study on children, adolescents and adults. In Moore, J. D. and Stenning, K., editors, *Proceedings of the XXIII Annual Conference of the Cognitive Science Society*, pages 875–880, Edinburgh, Scotland. Lawrence Erlbaum Associates.

Sánchez Valenzia, V., van der Wouden, T., and Zwarts, F. (1993). Polarity, veridicality and temporal connectives. In Dekker, P. and Stokhof, M., editors, *Proceedings of 9th Amsterdam Colloquium*, volume III, University of Amsterdam.

Sanders, T. J. M. and Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in processing. *Discourse processes*, 29(1):37–60.

Sanders, T. J. M., Spooren, W. P. M., and Noordman, L. G. M. (1992). Towards a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.

Sarkar, A. (2000). Practical experiments in parsing using tree adjoining grammars. In *Proceedings of the Fifth Workshop on Tree Adjoining Grammars (TAG+ 5)*, Paris, France.

Schilder, F. (2000). Robust text analysis via underspecification. In *Proceedings of Workshop ROMAND*.

Schilder, F. (2002). Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering*, 8(2–3):235–255.

Schilder, F. and Tenbrink, T. (2002). The interplay of information structure and the placement of 'after' and 'before'. In *Proceedings of the Workshop on Information Structure in Context*, Stuttgart, Germany.

Schulte im Walde, S. (2003). *Experiments on the Automatic Induction of German Semantic Verb Classes*. PhD thesis, Institut fūr Maschinelle Sprachverarbeitung. Published as AIMS Report 9(2).

Scott, D. and de Souza, C. (1990). Getting the message across in RST-based text generation. In Dale, R., Mellish, C., and Zock, M., editors, *Current Research in Natural Language Generation*. Academic Press.

Siddharthan, A. (2003). Preserving discourse structure when simplifying text. In *Proceedings of the 2003 European Natural Language Generation Workshop*, pages 103–110, Budapest, Hungary.

Siddharthan, A. (2005). Syntactic simplification and text cohesion. *Journal of Language and Computation*.

Siegel, E. V. and McKeown, K. (2000). Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–627.

Snow, R., Jurafsky, D., and Ng, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of the Conference on Neural Information Processing Systems*.

Sorace, A. and Keller, F. (2005). Gradience in linguistic data. *Lingua*, 115(11):1497–1524.

Spooren, W. (1989). *Some aspects of the form and interpretation of global contrastive coherence relations*. PhD thesis, Catholic University of Nijmegen, the Netherlands.

Sporleder, C. (2004a). Combining machine learning and set-theory to infer inheritance hierarchies. In *Proceedings of KONVENS-04*, Vienna, Austria.

Sporleder, C. (2004b). *Discovering Lexical Generalisations. A Supervised Machine Learning Approach to Inheritance Hierarchy Construction*. PhD thesis, School of Informatics, University of Edinburgh.

Sporleder, C. and Lapata, M. (2004). Automatic paragraph identification: A study across languages and domains. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain.

Sporleder, C. and Lascarides, A. (2004). Combining hierarchical clustering and machine learning to predict high-level discourse structure. In *Proceedings of The 20th International Conference on Computational Linguistics (COLING-04)*, Geneva, Switzerland.

Stede, M. (1994). A contrastive analysis of some contrastive discourse markers. In Quantz, J. and Schmitz, B., editors, *Ambiguity and Strategies of Disambiguation (Pro-ceedings of a Workshop held at the Annual Conference of the Deutsche Gesellschaft fur Sprachwissenschaft)*. Also available as Technical Report KIT-120,FB Informatik,TU Berlin,194.

Stede, M. and Umbach, C. (1998). DiMLex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pages 1238–1242, Montreal, Canada.

Sweetser, E. (1990). *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press, Cambridge, UK.

Tenbrink, T. and Schilder, F. (2001). (Non-)temporal concepts conveyed by 'before', 'after' and 'then' in dialogue. In *Proceedings of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue (BIDIALOG 2001)*.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles — experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4).

Thanh, H. L., Abeysinghe, G., and Huyck, C. (2004). Automated discourse segmentation by syntactic information and cue phrases. In *In Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2004)*, Innsbruck, Austria.

Tolkien, J. R. R. (1954). *The Lord of The Rings*. George Allen and Unwin, London.

Townsend, D. J. (1983). Thematic processing in sentences and texts. *Cognition*, 13:223–261.

van Dijk, T. A. (1979). Pragmatic connectives. *Journal of Pragmatics*, 3:447–456.

Vander Linden, K. (1994). Generating precondition expressions in instructional text. In *Proceedings of the 15th Conference on Computational Linguistics*, Kyoto, Japan.

Varges, S. and Mellish, C. (2001). Instance-based natural language generation. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*, pages 1–8, Pittsburgh, USA.

Vijay-Shanker, K. and Schabes, Y. (1992). Structure sharing in lexicalised tree adjoining grammar. In *Proceedings of COLING 92*, pages 205–211.

Villavicencio, A. (2001). *The Acquisition of a Unification-Based Generalised Categorial Grammar*. PhD thesis, Computer Laboratory, University of Cambridge.

Webber, B. and Joshi, A. (1998). Anchoring a lexicalized tree-adjoining grammar for discourse. In *ACL/COLING Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada.

Webber, B., Stone, M., Joshi, A., and Knott, A. (1999). Discourse relations: A structural and presuppositional account using lexicalised TAG. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, Maryland, Canada.

Webber, B., Stone, M., Joshi, A., and Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–588.

Webber, B. L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6:107–135.

Weeds, J. (2002). Asymmetry in a similarity measure. In Weeds, J., editor, *The 15th Whitehouse Papers*. University of Sussex.

Weeds, J. (2003). *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, University of Sussex.

Weeds, J. and Weir, D. (2003). A general framework for distributional similarity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, Sapporo, Japan.

Weeds, J., Weir, D., and McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference of Computational Linguistics (COLING-2004)*, Geneva, Switzerland.

Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of HLT/NAACL 2003*, pages 276–283.

Wilks, Y. (2004). On the ownership of text. *Computers and the Humanities*, 38(2):115–127.

Williams, E. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society (Series B)*, 21:396–399.

Williams, S. (2004). *Natural Language Generation (NLG) of discourse relations for different reading levels*. PhD thesis, Department of Computing Science.

Williams, S., Reiter, E., and Osman, L. (2003). Experiments with discourse-level choices and readability. In *Proceedings of the 9th European Workshop on Natural Language Generation*, pages 127–134, Budapest.

Witten, I. H. and Frank, E. (2000). *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco.