An Information Theoretic Approach to Production and Comprehension of Discourse Markers

Thesis for obtaining the title of
Doctor of Engineering Science
of the Faculty of Natural Science and Technology I
of Saarland University

by

Fatemeh Torabi Asr
Saarbruecken
October 2015

Day of Colloquium
10th of November 2015

Dean of faculty
Prof. Dr. Markus Bläser

Chair of comittee
Prof. Dr. Manfred Pinkal

First reviewer
Dr. Vera Demberg

Second reviewer
Prof. Dr. Antonio Krüger

Third reviewer
Prof. Dr. Ivan Titov

Academic assistant
Dr. Asad Sayeed

Abstract

Discourse relations are the building blocks of a coherent text. The most important linguistic elements for constructing these relations are discourse markers. The presence of a discourse marker between two discourse segments provides information on the inferences that need to be made for interpretation of the two segments as a whole (e.g., *because* marks a reason).

This thesis presents a new framework for studying human communication at the level of discourse by adapting ideas from information theory. A discourse marker is viewed as a symbol with a measurable amount of relational information. This information is communicated by the writer of a text to guide the reader towards the right semantic decoding. To examine the information theoretic account of discourse markers, we conduct empirical corpus-based investigations, offline crowd-sourced studies and online laboratory experiments. The thesis contributes to computational linguistics by proposing a quantitative meaning representation for discourse markers and showing its advantages over the classic descriptive approaches. For the first time, we show that readers are very sensitive to the fine-grained information encoded in a discourse marker obtained from its natural usage and that writers use explicit marking for less expected relations in terms of linguistic and cognitive predictability. These findings open new directions for implementation of advanced natural language processing systems.

Zusammenfassung

Diskursrelationen sind die Bausteine eines kohärenten Texts. Die wichtigsten sprachlichen Elemente für die Konstruktion dieser Relationen sind Diskursmarker. Das Vorhandensein eines Diskursmarkers zwischen zwei Diskurssegmenten liefert Informationen über die Inferenzen, die für die Interpretation der beiden Segmente als Ganzes getroffen werden mssen (zB. *weil* markiert einen Grund).

Diese Dissertation bietet ein neues Framework für die Untersuchung menschlicher Kommunikation auf der Ebene von Diskursrelationen durch Anpassung von Ideen aus der Informationstheorie. Ein Diskursmarker wird als ein Symbol mit einer messbaren Menge relationaler Information betrachtet. Diese Information wird vom Autoren eines Texts kommuniziert, um den Leser zur richtigen semantischen Decodierung zu führen. Um die informationstheoretische Beschreibung von Diskursmarkern zu untersuchen, führen wir empirische korpusbasierte Untersuchungen durch: offline Crowdsourcing-Studien und online Labor-Experimente. Die Dissertation trägt zur Computerlinguistik bei, indem sie eine quantitative Bedeutungs-Repräsentation zu Diskursmarkern vorschlägt und ihre Vorteile gegenüber den klassischen deskriptiven Ansätzen aufzeigt. Wir zeigen zum ersten Mal, dass Leser sensitiv für feinkörnige Informationen sind, die durch Diskursmarker kodiert werden, und dass Textproduzenten Relationen, die sowohl auf linguistischer Ebene als auch kognitiv weniger vorhersagbar sind, häufiger explizit markieren. Diese Erkenntnisse eröffnen neue Richtungen für die Implementierung fortschrittlicher Systeme der Verarbeitung natürlicher Sprache.

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Vera Demberg for the continuous support of my PhD study, her flexibility in accepting my topics of interest and her dedication to our meetings and discussions. Her guidance helped me in designing the experiments, presentation of the work to different communities and writing of this thesis. Besides my supervisor, I would like to thank the rest of my thesis committee. Prof. Dr. Antonio Krueger and Prof. Dr. Ivan Titov reviewed the desertion, and Prof. Dr. Manfred Pinkal honored me by chairing the colloquium. Their insightful comments and questions widened my view for the future research. I had the pleasure of being officemates with Dr. Asad Sayeed, the most fun and knowledgeable colleague ever, who was also in my defense committee. During the past four years, we discussed a variety of topics in Computer Science, AI and Linguistics through which I learned a lot from him.

My sincere thanks also goes to the following professors for providing insightful comments, organizing my research visits and providing access to laboratory facilities: Prof. Dr. Matthew Crocker, Prof. Dr. Bonnie Webber, Prof. Dr. Florian Jaeger, Prof. Dr. Ted Sanders, Prof. Dr. Frank Keller, Prof. Dr. Hannah Rohde, and Prof. Dr. Manfred Stede.

My colleagues in Saarbruecken and fellow scientists that I met in Potsdam, Edinburgh and Utrecht during short research visits provided very useful feedback on different portions of the research. I would also like to thank the audience of my talks and presentations at all the relevant venues, including the AMLaP, CUNY, ACL, CMCL and Coling conferences.

I am grateful to the two institutes at Saarland University who funded this research: the MMCI, Cluster of Excellence on "Multimodal Computing and Interaction" and the Collaborative Research Center SFB 1102 on "Information Density and Linguistic Encoding" (under German Research Foundation).

Warm acknowledgement goes to the friends I met during my PhD. They made these years a unique, adventurous and memorable experience. Annemarie Friedrich, David Howcroft, and Merel Scholman helped considerably with proof reading of this manuscript. Thanks to Elli Tourtouri, Alessandra Zarcone, Ines Rehbein, William Blaco, Ekaterina Kravtchenko, Jorrig Vogels and many others for the friendly and vibrant environment they created at office. Thanks to Rahil Mahdian and his family, Maryam Nazarieh, Fatemeh Behjat, Sara Pahlavan, Hosna Sattar, Javad Najafi, Amirhosein Kardoost, Majid Azimi, Fahimeh Ramezani, Ali Pourmiri and all other Iranian friends who made Saarbruecken feel more like home than it would have been without them. Thanks to my Persian course

students at the language center of Saarland University for their shared interest to gather and learn about different cultures. The last five years of my life have been as influential and enjoying as the preceding twenty five years altogether because of these people.

Last but not the least, I would like to thank my family: my parents and my brothers for supporting me spiritually throughout my life. My father has always been the source of inspiration and example of perseverance at work. My mother has always provided a great moral support in hardships and reminded me of the big picture of life. I cannot imagine pursuing my academic career without their everlasting love and encouragement.

For the one who sends the snowflakes,

... and for my grandparents.

Table of Contents

Chapter 1

Introduction

1.1 The theoretical framework

Human language processing can be viewed from three perspectives: acquisition, comprehension and production. Each of these interrelated processes can be investigated at different conceptual levels that are traditionally categorized into phonology, morphology, syntax, semantics and pragmatics. This study is about comprehension and production of discourse relations dealing with the semantic and pragmatic levels. Previous information theoretic studies of language and human communication provide evidence for systematic interactions between comprehension and production processes. However, the majority of work in this domain has focused on shallower levels of sentence processing involved with phonology, morphology, syntax and sentence-internal semantics. An information theoretic account of human communication views linguistic elements such as words and phrases as units of information being transferred from a speaker to a listener. We propose that discourse-level language processing can also be explained in such a framework. Discourse markers such as sentence connectives are the most important triggers for establishing relations between sentences in a text or utterance.[1] Thus we focus on these elements to begin developing a new framework for studying discourse-level comprehension and production and the interaction between the two, in a similar vein to what has been done for other levels of human sentence processing.

1.2 Research questions

While discourse relations and discourse connectives have been very widely studied in both theoretical linguistics and natural language processing, we still do not have a unified and quantified account that can tell us how the two phenomena are related to one

[1]The terms discourse/sentence connective and discourse marker are used interchangeably in this thesis, unless we explicitly distinguish between the two.

another. Studying them within an information theoretic framework will let us answer the following questions:

- What type and amount of relational information is delivered by a discourse connective (e.g., "but" or "although")?
- How does this information affect comprehension and production of discourse relations?

These are important theoretical questions in the domain of semantics and pragmatics that fall beyond the scope of classic grammatical approaches to language. Once we know the answer to these questions, solving a lot of practical problems in multi-sentence text processing (e.g., machine translation, text summarization, question answering, etc.) will also become easier. For example, once we can define the type of context in which a discourse connective would be required vs. redundant, then we can design more natural sounding language generation systems.

1.3 Methodology

The first thing we need to address the above questions is a method of calculating the information content of a discourse connective. We propose that this can be done by collecting occurrences of a discourse connective from a corpus of natural text that is annotated with discourse relations between sentences. The distribution of the discourse relations a connective co-occurs with will be used as a representation of its information content. This gives a straightforward answer to the first question under investigation, but to make sure that this method of calculation is psycholinguistically plausible, we conduct a set of experiments. A crowd-sourced survey study investigates the differential effects of similar connective types on coherence of a text. It is followed by an eye-tracking reading-time experiment to measure readers' sensitivity to the effect of discourse connectives during online reading. Results of these experiments indicate that even very fine grained differences between two connectives, in terms of how they are distributed across relations of different types in production data, show up in comprehension. More specifically, keeping the context identical, each connective type has a unique effect on interpretation of the relation in which it is utilized and this effect can be predicted by its information content calculated based on corpus data. This first set of experiments shows how production data can predict the way a discourse connective is comprehended.

In order to answer the second question from a production point of view, we take the

opposite methodological direction: predicting production patterns based on comprehension experiments. First, we identify other sources of relational information and cognitive biases of listeners for discourse-level interpretations. Recent experimental work on incremental sentence processing, cognitive theories of narration interpretation, and previous theories on information and communication are all put together to formulate a set of hypotheses regarding the natural use of discourse connectives. In line with the predictions of the information theoretic account and previous findings regarding other levels of production, we find that discourse connectives tend to appear in contexts where the relations they mark are less predictable and they tend to be dropped by speakers when the relation is predictable given its context. This is consistent with the hypothesis that speakers formulate their utterances in a way that is informed by a listener model, i.e., selection of the form used to deliver a message involves the consideration of comprehension-side constraints. The results of all experiments in the thesis put together corroborate an account of human discourse processing which involves a strong interaction between comprehension and production behaviors is involved.

1.4 Contributions

In addition to a comprehensive review of the previous work in NLP and psycholinguistics on discourse relations and their markers, the following contributions are made to the field:

- We propose an information theoretic representation of discourse markers, and via that, we address a set of questions about the ambiguities associated with discourse connectives and inferences in under-specified contexts, which the classic descriptive approaches have left unanswered.
- We show that even very fine-grained information encoded in a discourse connective can be approximated by looking into large corpus of natural text and this information influences the offline and online discourse comprehension processes.
- We raise the issue of connective reduction as a language production behavior and provide evidence that the decision of the writer for using or dropping discourse connectives is sensitive to a set of linguistic and non-linguistic factors.
- We examine a set of general to specific theories on human communication at the level of discourse in an empirical setting. These include the Uniform Information

Density theory (Levy and Jaeger, 2007), the continuity hypothesis (Segal et al., 1991; Murray, 1997) and the causality-by-default hypothesis (Sanders et al., 1992).

Finally, several research directions are proposed for future work on automatic identification of discourse relations and development of more natural sounding language generation systems.

1.5 Overview of the chapters

Chapter 2 provides an introduction to discourse relations and discourse connectives. The importance of discourse relations as the basis for discourse analysis will be discussed in this chapter. Furthermore, we look into available resources for a computational study, and in particular calculation of a discourse connective's information content. Other linguistic markers of discourse relations are also reviewed in the context of current machine learning attempt to automatically identify discourse relations. Some of these features will be referred to later in Chapter 5 when we investigate the question of connective use necessity/redundancy.

Chapter 3 starts with an overview of previous information theoretic studies at other levels of sentence processing which motivate the general proposal of the thesis. Then, a method is presented for calculation of connective information content and investigating discourse processing from an information theoretical perspective. The final part of the chapter presents an overall analysis of the connectives and relations in Penn Discourse Treebank. Our more focused research questions are shaped throughout this exploratory study:

- **Specificity of a connective to a relation:** what type of relational information is encoded in different discourse connectives? what is the granularity of the relation a connective marks?
- **Ambiguous and multi-sense connectives:** what types of ambiguity are involved with the meaning of a discourse connective? what are multi-sense connectives? what can we predict about these connectives by looking into their natural usage?
- **Implicitness of the discourse relations:** what types of relations can be expressed with or without discourse connectives? how often are connectives dropped in natural text and why?

Chapter 4 includes our first set of experiments, which investigates discourse **compre-**

hension. A case study on *but* and *although* is conducted to see if the distribution of these connectives in naturally occurring text can tell us about the way they influence the comprehension of short stories. Two offline and one online reading experiments show that even when both connectives fit equally well in a certain context they generate different inferences, thus different expectations regarding the way the discourse will be continued. All differences that we observe between *but* and *although* correlate with the way the two connectives are distributed across discourse relations of different types. This corroborates the reliability of the information content calculation based on discourse annotated corpora, and more importantly, it provides evidence for the interaction between discourse comprehension effect of a connective and its production patterns.

Chapter 5 includes the second set of our experiments focused on the question of implicitness; That is, when discourse connectives are utilized in **production** for marking relations explicitly and when they are reduced. Previous theories about discourse comprehension emphasize that readers expect specific types of relation between sentences such as cause-consequence and continuous temporal relations and they experience processing difficulty when these expectations are not met. We also know (based on the studies reviewed in Chapter 2) that discourse relations can be identified not only by connectives but also by other linguistic features of the involved sentences. Based on this data, we hypothesize that a reason for dropping discourse connectives could be avoiding redundancy. we perform a high-coverage analysis of discourse relations in PDTB to validate this hypothesis. We find that discourse connectives tend to be present more often when the relation they mark is unexpected (discontinous and non-causal relations), or when other linguistic cues of the relation sense they mark are absent. On the other hand, connectives of generally expected relations (causal and continuous) or relations that are marked by other linguistic means tend to be dropped by the speakers. This provides additional evidence for a systematic relation between comprehension and production mechanisms at the level of discourse relations.

Chapter 2

Background

This chapter presents an overview of the linguistic research on text coherence and discourse relations with an emphasis on the computational approaches. I will start by a brief introduction of the major theories of discourse coherence in the field. Discourse relations and the way they are categorized particularly in Penn Discourse Treebank, the main resource for the corpus-based chunk of my study, are introduced. Finally, the linguistic markers of discourse relations ranging from sentence connectives to syntactic, semantic and clause-level features of the relational arguments that have been used for identification of relations in expository text will be reviewed. This background information is essential, as I refer to the introduced terms, concepts and theories in the rest of the thesis. It also gives a chronological view of the progress in discourse relation research and familiarizes the reader with the state-of-the-art machine-learning approaches to the problem.

2.1 Introduction to discourse relations

According to the work of linguists over decades, we now can describe to good extent what a grammatical sentence looks like. In English, presence of a verb and certain arguments, the order they appear one after the other and the message it all together delivers about the external world explain whether the composition is an acceptable sentence or just a random sequence of words. In a relatively similar way that the parts of speech, dependencies, tree structures and semantic roles within a sentence's territory have been developed, Halliday and Hasan (1976) attempted to define a grammar for discourse. They identified several cohesive devices in multi-sentence text that should be utilized to shape a sensible discourse: reference, substitution, ellipsis, conjunction and lexical chains. Despite of it being a prominent work still referred frequently by the researchers in the field, Halliday and Hasan's theory has been criticized because of its grammatical approach to the analysis of discourse. Perhaps with the exception of lexical chains (which refer to semantic relations between words), all other devices pertain to the very surface

characteristics of a text. For example, anaphoric pronouns establish hardware connections between neighboring sentences but do not guaranty that the composition makes sense (see a clarifying example in (1), where the mere use of anaphora is not sufficient for a coherent combination). Coherence is the holistic property of a well-formed discourse and is achieved at the semantic and pragmatic levels. An extended discussion on this debate can be found in Carrell (1982).

(1) a. Amy wanted the doll. **It** was so cute.
b. ?Amy wanted the doll. **It** was so ugly.[1]

A few years later, Hobbs (1979) proposed an alternative approach to define and investigate text coherence that was based upon ideas from logic and propositional inference. Hobbs focuses on the semantic relations between sentences and how they can be inferred by human brain or automatic systems. From this perspective, a coherent text would be one composed of sentences that are connected to one another via inferential relations, while cohesive devices might just contribute to construction of these relations (Sanders et al., 1992; Sanders and Noordman, 2000). Several theories of discourse coherence have been developed since then that, in one way or another, deal with discourse relations. In the following, I will introduce a selection of these theories that successfully found broad audience in the linguistic community, became reference models for annotation of large-scaled discourse corpora, and finally, established new research directions for discourse analysis in theoretical and application oriented settings.

The theory of coherence and coreference by Hobbs (1979) defines an inference system based on four components including *data*, *representation*, *operations* and *control*. Data corresponds to the worldknowledge that the system has access to in some representation. For example, a native speaker of English has access to a set of commonly possessed knowledge in the shape of axioms. The important components in an inference system are the operations that can be applied to data in certain ways specified by controls. When a clause in a text is encountered, at least one proposition is asserted and all axioms whose antecedents are satisfied according to the asserted proposition can be activated, i.e., some inferences will be made. Discourse relations are formulated in a pseudo-predicate logic format and are explained in detail with the help of concrete examples from natural language. For example, the `Elaboration` relation is defined as follows:

[1]Throughout the thesis, I indicate the ungrammatical or incoherent variants of an example text by putting a question mark in front of it.

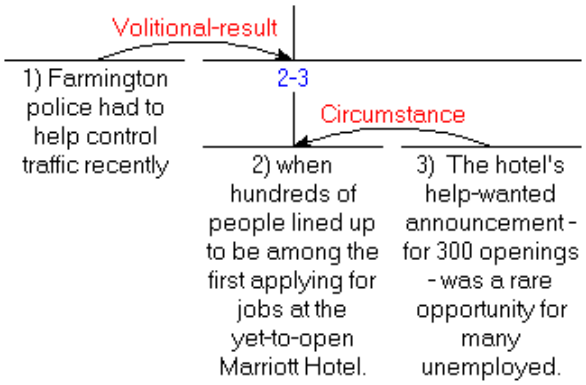


Figure 2.1: A sample text analyzed by RST relations (Carlson et al., 2003)

(2) S1 is an `Elaboration` of S0 if a proposition P follows from the assertions of both S0 and S1 (but S1 contains a property of one of the elements of P that is not in S0).
Example: *Go down Washington Street. Just follow Washington Street three blocks to Adams Street.*

Hobbs' theory also explains how inferring semantic relations between propositions underlying sentences interacts with other discourse-level processes. This continues to be an interesting research topic, e.g., recent experimental studies confirm the explanatory power of a coherence-based approach over other accounts of pronoun resolution which emphasize on grammatical aspects (Kertz et al., 2006; Kehler et al., 2008; Kaiser, 2009; Rohde and Kehler, 2014). The set of discourse relations originally formulated by Hobbs does not cover the variety found in coherent natural text, thus a lot of following studies by other researchers have been conducted on defining new relation senses. We get back to this point in Section 2.2.

The rhetorical structure theory (RST) by Mann and Thompson (1988) approaches discourse coherence from a text analyst perspective. It explains in what way the reader is supposed to organize different pieces of a text in order to construct a coherent representation of the underlying story. The RST analysis of a text starts off by looking up relations between adjacent clauses, which are combined and in turn considered as larger discourse segments. This process is continued until a rhetorical tree covering the entire text is obtained (see Figure 2.1). A relation is composed of more important vs. less important spans called *nuclei* and *satellite*, respectively. Each relation type is defined with respect to the constraints put on each of its arguments, as well as an intended effect on the reader.

For example, the definition of an `Evidence` relation is as follows:

(3) Constraint on Nuclei: reader might not believe in N to a degree satisfactory to writer.
Constraint on Satellite: the reader believes S or will find it credible.
Constraint on the combination: reader's comprehending S increases reader's belief of N.
Example: *The program as published for calendar year 1980 really works. In only a few minutes, I entered all figures from my 1980 tax return and got a result which agreed with my hand calculations to the penny.*

The RST account has been very popular in application-oriented research (a comprehensive review has been done by Taboada (2006)). A variety of semantic and pragmatic discourse relations are defined and utilized for text analysis in this literature depending on the requirements of the specific applications. The notion of nuclearity distinguishes RST relation s from most other discourse relation categorization systems. Finding the more or less important pieces of information in a multi-sentence discourse becomes possible via nuclearity detection, but its criteria have been a matter of controversy. We will talk more about this dimension of a discourse relation in the following chapters. The success of RST in improving performance of a variety of automated language processing systems proves the multi-faceted power of an account of coherence based on discourse relations.

The discourse structure theory by Grosz and Sidner (1986) identifies three distinct but interacting discourse structures. The *linguistic* structure is composed of the discourse segments that are connected to one another via surface markers such as discourse connectives and cue phrases without forcing a particular system of semantic relations to be assigned as part of the analysis. The *intentional* structure demonstrates how the purpose behind a discourse segment is related to the purpose of another. They identify two relations at this level, the *dominance* vs. the *satisfaction-precedence* between discourse segments, which mimics the *nuclei*/*satellite* distinction in RST. Finally, the *attentional* or focusing structure determines the salience of the discourse entities in the context and how it changes throughout the discourse. The idea of a multi-layer discourse analysis in this theory brings attention to the very complicated structure of discourse and possibilities to decompose the problem into simpler sub-problems. For example, it introduces new directions for studying information structure (Grosz et al., 1995; Lambrecht, 1996). Grosz and Sidner's approach differs from Hobbs' that focuses on the local semantic inferences but still involves a notion

of discourse relation at the intentional level. In general, Grosz and Sidner's theory shares more commonalities with the RST (see Moser and Moore (1996) for a discussion) than with other theories of discourse. In terms of the computational competence, it remains more abstract and less applied to natural language processing systems.

The segmented discourse representation theory (SDRT) by Asher and Lascarides (2003) is an extensively worked out logic which tries to combine notions of formal semantics at the sentence level with those of discourse structure and rhetorical relations. The main difference between this account with that of Hobbs is that SDRT distinguishes between the logical representation of an utterance and its interpretation. Representations follow a stricter logical form and operations are applied in a systematic manner. For example, remember the `Elaboration` relation defined by Hobbs (1979); This relation in an SDRT analysis of text is represented in the following way:

(4) $\texttt{Elaboration}(\alpha, \beta) \vdash\Downarrow (\alpha, \beta)$

and is added to a system of predicates as a consequence of the following discourse update:

(5) $(?(\alpha, \beta, \lambda) \wedge \mathrm{TOP}(\sigma, \alpha) \wedge \mathrm{subtype}_D(\sigma, \beta, \alpha) \wedge \mathrm{Aspect}(\alpha, \beta))$
$> \mathrm{Elaboration}(\alpha, \beta, \lambda)$

where TOP, $\mathrm{subtype}$ and Aspect each represents another active predicate in the system. $\texttt{Elaboration}(\alpha, \beta)$ would then add the following predicate to the system (this is refereed to as a *temporal consequence* of the `Elaboration` relation):

(6) $\phi_{Elaboration(\alpha,\beta)} \Rightarrow \mathrm{Partof}(e_\beta, e_\alpha)$

which means the event α should be considered as part of the event β. Compare this relation with `Explaination` that is the result of a slightly different discourse update:

(7) $(?(\alpha, \beta, \lambda) \wedge \mathrm{TOP}(\sigma, \alpha) \wedge \mathrm{cause}_D(\sigma, \beta, \alpha) \wedge \mathrm{Aspect}(\alpha, \beta))$
$> \mathrm{Explanation}(\alpha, \beta, \lambda)$

and has different temporal consequences in the system:

(8) a. $\phi_{Elaboration(\alpha,\beta)} \Rightarrow (\neg e_\alpha \succ e_\beta)$

b. $\phi_{Elaboration(\alpha,\beta)} \Rightarrow (\text{event}(e_\beta) \Rightarrow e_\beta \succ e_\alpha)$

In SDRT, interpretations emerge as a result of *discourse coherence maximization* principle. Degree of coherence of a candidate interpretation is measured based on a set of rules, such as "the more anaphoric expressions whose antecedents are resolved, the higher the quality of coherence of the interpretation." or "the more rhetorical connections there are between two items in a discourse, the more coherent the interpretation". SDRT is one of the most worked out computational frameworks for analyzing discourse operations in connection with propositional-level semantics, which explains itself in the context of theoretically important discourse phenomena such as anaphora and presupposition. However, the huge glossary of its formal definitions needs more proficiency to work with and that might be one reason why after more than a decade of its birth this account has not taken the place of the less formal approaches in NLP or theoretical research.

The common idea behind the theories I just reviewed is that semantic analysis of a discourse by a machine or a human is based upon the identification of relations between neighboring clauses and sentences. This indicates that in order to construct an automatic system capable of discourse-level processing of natural language or to model human discourse comprehension as a computational process for psycholinguistic purposes, we need to familiarize ourselves with discourse relations. The following sections provide insights into some theoretical and practical problems involved with discourse relation annotation and identification.

2.2 Categories of discourse relations

The above reviewed theories constitute the most prominent approavhes in computational linguistics that rely on clause-level relations to explain discourse phenomena. Relations of this kind have in fact a broader history with some roots in the work of philosophers such as Aristotle and Hume (1784). Unfortunately, not much consensus has been obtained to date in terms of the number and type of relation senses that should be considered in a standard discourse analysis. For example, even in studies following the Rhetorical Structure Theory several practices of relation sense categorization have been

adapted.

In 1992, Hovy and Maier reviewed the work of 25 scientist from different disciplines, including linguistics, computational linguistics and psycholinguistics, and compiled an agreement taxonomy including 16 distinct mid-grained categories of rhetorical/semantic relations (Fig. 2.2). They also proposed that the intentional structure should be considered orthogonal to the semantic interconnections, i.e., inferences can be analyzed by factoring out the intentional aspects. Nevertheless, later studies often introduce their own taxonomy or categorization of discourse relations (Sanders et al., 1992; Longacre, 1996; Kehler, 2002). Inconsistencies in discourse relation categorization systems not only results in a difficult scientific communication among researchers working on similar topics, but in the first place questions the plausibility of an inference-based account of discourse coherence. This challenge is also reflected in recent efforts for achieving a standard system by the ISO community (Bunt et al., 2012) and the European mission for structuring discourse (TextLink, 2015). One proposal is to work on mappings between pairs of relation categorization systems. A recent work in this direction is a schema to map between RST and SDRT relations proposed by Zitoune and Taboada (2015). Another idea is to develop a unifying framework for mapping between any two relation categorization systems based on a set of cognitively motivated dimensions (Sanders et al., 2016). These dimensions were first motivated by Sanders et al. (1992) and are representative of the basic features the human mind is able to extract from a relation between two sentences. For example, an `Explanation` relation as defined in RST or SDRT involves causality, whereas `Elaboration` does not. If basic dimensions as such can be identified for relations of different types then a mapping becomes possible between any two sets of discourse relations. The application-oriented research can employ an arbitrary system of discourse relations as long as it works for a specific task, but mappings need to be performed in cross-system, cross-language, or cross-corpora studies. Even more concern is involved with developing computational models of human discourse processing and studying language as a natural phenomenon. Given the discrepancy of the relation categorization systems across theories of discourse inference, the question is: would it be *possible* and *plausible* to test the validity of a cognitive hypothesis via computational modeling if a model is constructed on the basis of an arbitrarily selected system of relation sense categorization?

Our experiments in chapters 4 and 5 provide a positive response to the question of *possibility* that I just pointed out. In Chapter 4 we look into the a corpus of text annotated based on one of the widely utilized systems of relation sense categorization (Penn Discourse Treebank, which I will shortly introduce) and extract co-occurrence of

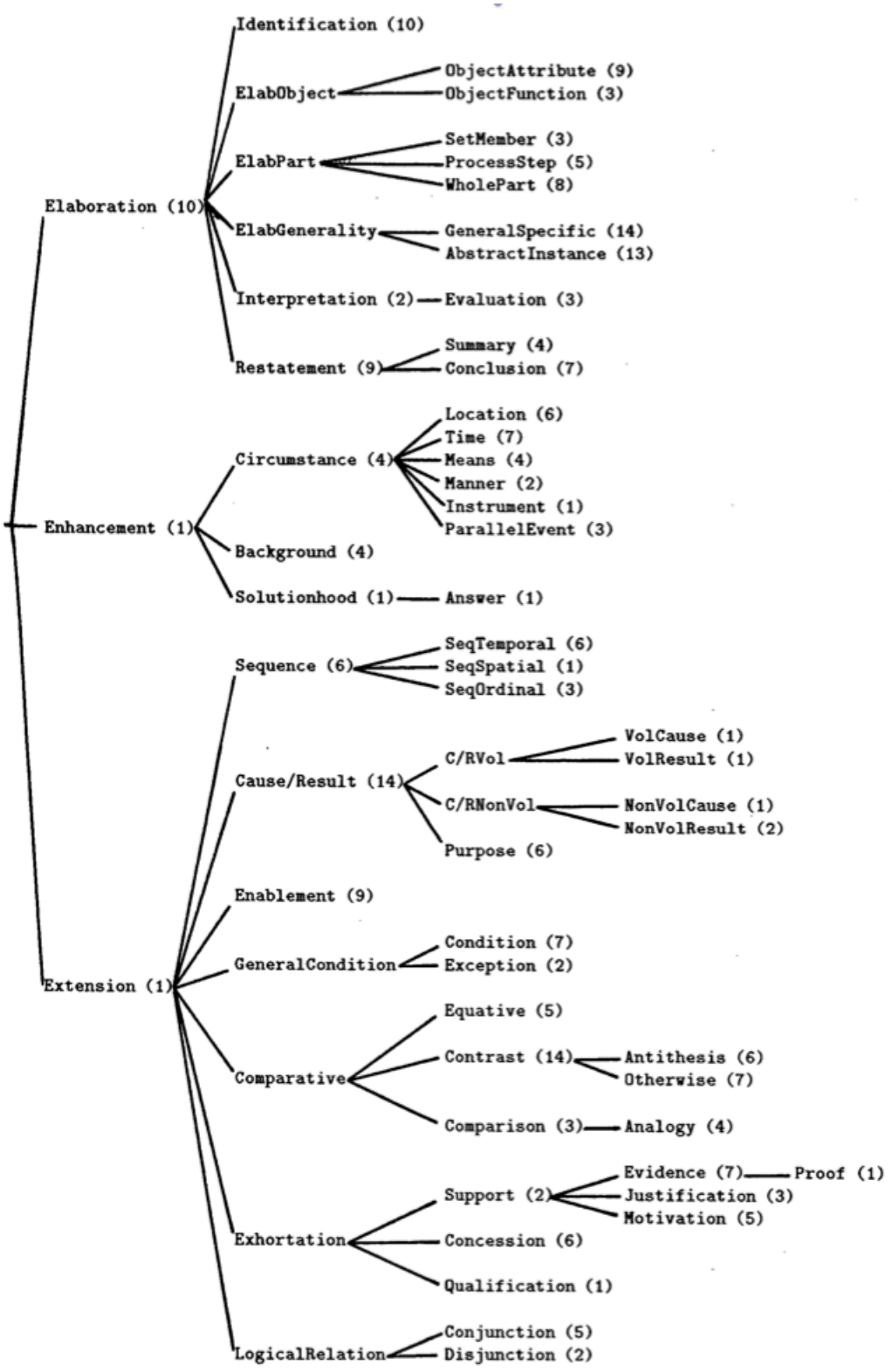


Figure 2.2: Discourse relations collected and merged by Hovy and Maier (1992).
The number in front of each relation indicates the number of researchers who proposed it.

those relation senses with discourse connectives of different types. We then predict the effect of discourse connectives on inferences based on these statistics and examine our predictions through crowd-sourced coherence judgment and eye-tracking experiments on human subjects. Chapter 5, takes an opposite direction: we hypothesize based on previous experimental studies what patterns we should find in a corpus of natural text and then examine these predictions by analysing PDTB relations. Results of both sets of experiments are encouraging, i.e., they provide some psycholinguistic support for the annotation system of PDTB to be sensible. The question of *plausibility*, however, needs far more research of this type. Agreement between the results of lab and computational experiments on a given phenomenon provides stronger evidence for a confirmed hypothesis. Yet one should be careful about the generalizations made on top of a finding as such. The next section provides an overview of the common annotation strategies and available corpora of discourse relations and the reasons why we chose PDTB for the corpus-based portion of our study.

2.3 Annotated resources

The most popular discourse-annotated corpora in English are Penn Discourse Treebank (Prasad et al., 2008), RST-DT (Carlson et al., 2003) and Discourse GraphBank (Wolf et al., 2005). Each of these resources has adapted a different annotation schema which might have been evolved and modified in the course of annotation. The following set of features distinguish Penn Discourse TreeBank from the other resources and makes it particularly suitable for our study:

- With about 50 k relation instances, PDTB is the biggest resource of discourse relations in size.
- The text comes from sections of Wall Street Journal that have been also manually annotated for syntax in the Penn Treebank project (Marcus et al., 1993).
- In addition to discourse relations, explicit and implicit (artificially inserted) discourse connectives are annotated in PDTB. Having gold-standard syntactic annotation and discourse connectives in place facilitates our analyses and adds to the reliability of the conclusions (this will become clear in later chapters of the thesis).[2]

[2]More recently, Taboada and Das (2013) have extended the annotation of the RST-DT corpus with discourse connectives as well as entity features, and a range of lexical, syntactic, graphical and numerical features. This corpus will soon become available and might be useful for a similar study on discourse markers.

- The PDTB schema of relation sense categorization has become very popular and been adapted for development of discourse corpora in other languages. Thus basing our hypothesis testing on this data source would provide a better basis for future comparative studies.

In addition to the above features, PDTB annotators have also achieved a relatively good level of inter-annotator agreement because they employed stricter and perhaps more objective instructions for detecting discourse segments and identifying relation senses. For example, in PDTB discourse units (relational arguments) are clauses or full sentences and the annotators are asked to find the minimal text span of this type to indicate relational arguments, whereas in RST-DT discourse segments or relational arguments can vary between a phrase and a paragraph. Also, for relation sense identification, PDTB uses a lexico-semantic strategy, that is to first find a discourse connective like *because* in the text (or insert one, if the arguments are independent discourse segments) and then assign a relation tag, e.g., `reason`. Annotators of Discourse GraphBank were also instructed to use the connective substitution strategy to find the relations but connectives are not part of the annotations. In RST-DT, discourse connectives are only used for determining the boundaries of the discourse segments, i.e., they do not have an official status in determining the type of discourse relation.

The co-annotation of the discourse relations and discourse markers is in fact a fundamental characteristic of PDTB when compared against other corpora. The idea comes from the proposal of Knott and colleagues who motivated a taxonomy of coherence relations based on the connective types utilized for expressing them (Knott and Dale, 1994; Knott, 1996; Knott and Sanders, 1998). This proposal comes with an empirical study of discourse connectives in English and Dutch. Native speakers of the languages were recruited to first identify discourse markers in a corpus of natural text and then examine them for substitutability. Some connectives like *but* turned out to be good substitutes for a wide range of different connective types (e.g., *however* and *nevertheless*), whereas others only filled in very specific relations. While this is a neat methodology for defining and annotating relations in natural text, it also has a few drawbacks: 1) the resulting taxonomy of relations will be language dependent, and 2) relation senses that have no explicit marker cannot be identified, or if they are identified with some extra work it would be difficult to find their place within the taxonomy.

Another simplification considered in annotation of PDTB relations compared to the other two corpora is involved with the notion of discourse hierarchy. While in RST arguments of a relation are distinguished regarding their weight or importance (the notion

of nuclearity that I introduced before) and in Graphbank relations are divided into directed and undirected types, PDTB relations only come with a canonical semantic definition. In Chapter 4, I explain which PDTB relation types encode some sort of asymmetry between the two arguments that is implicit in the definition of the relation senses.

In PDTB, annotation of relations is considered for pairs of clauses connected by a discourse connective, as well as between neighboring sentences which are not connected by any discourse cue. From a syntactic and structural point of view, PDTB annotation is grounded in the framework of a Lexicalized Tree-Adjoining Grammar for Discourse (DL-TAG Webber et al.). All relations are composed of two discourse segments which are called Arg1 and Arg2. In *explicit* relations Arg2 is the segments that is syntactically attached to the discourse connectives, and in other relations, it refers to the argument appearing later in text. If two clauses are joined by a discourse connective, the boundaries of the arguments are annotated and a label indicating the relation sense is assigned. In this case the relation is categorized as explicit. To find the connectives, PDTB annotators started their task with a list of discourse connectives, but this list has been expanded in the course of annotation. For unconnected neighboring sentences and clauses, the annotators were asked to first see whether any discourse connective could artificially be inserted between the two arguments. These connectives are made available as part of the annotation along with the relation sense they are intended to mark. In this case the relation belongs to the *implicit* category, since the connective is not part of the original text. Unlike RST annotation which continues by assigning higher level relations to extended spans of text and constructing a tree-like structure of the entire text, PDTB relations are only annotated for very locally related sentences and clauses. In the course of annotation, wherever the annotators found unconnected neigbouring sentences but did not manage to insert a connective, they looked for *Alternative Lexicalization* of a relation. These are expressions such as *that is why* which encode specific relations but are not traditionally considered as discourse connectives. These relations are also annotated in the corpus under the category of *AltLex*, rather than *explicit* or *implicit*. If a given pair of neighboring discourse segments does not fit into one of the three mentioned categories, the annotators examine whether any common entity is mentioned in the two sentences. If yes, it will be tagged as an *EntRel* relation, and otherwise as a *NoRel*.

PDTB relation senses that the annotators used for labeling implicit, explicit and AltLex relations are organized in a hierarchy of coarse- to fine-grained categories depicted in Figure 2.3. This hierarchy is designed by consideration of the previous work on discourse

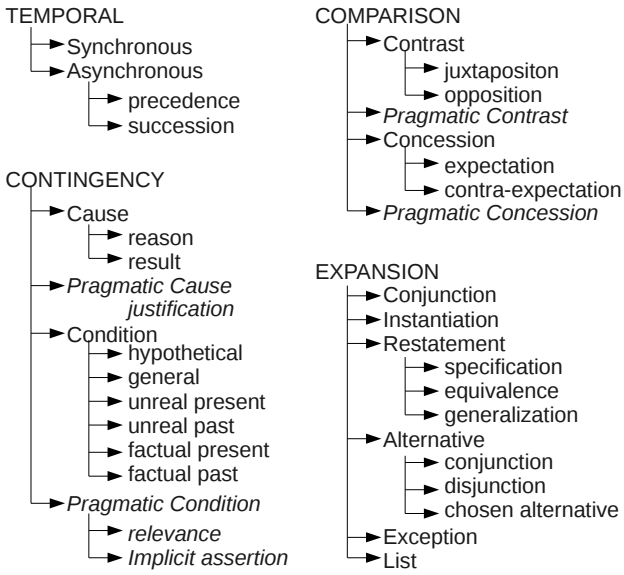


Figure 2.3: Hierarchy of relation senses in PDTB (Prasad et al., 2008)

relation taxonomies and particular inspirations from Hobbs (1979) and Knott (1996).[3]

To give an example, the explicit relation in (9) appears in the text with connective *instead* and is tagged as a `chosen alternative` relation, a very specific label located in depth three of the hierarchy. This label applies when the two arguments of the relation, i.e., Arg1 and Arg2 denote alternative situations but only the one stated in Arg2 has occurred. Throughout the following chapters I refer to the definitions and examples of other PDTB relations whenever necessary.

(9) *[No price for the new shares has been set.] Instead, [the companies will leave it up to the marketplace to decide.]* — `EXPANSION.Alternative.chosen alternative`

According to Miltsakaki et al. (2008) some relations are annotated with less specificity due to the disagreement between annotators, or with two different sense labels when both relations are conveyed simultaneously. Inter-annotator agreement has been reported as 94%, 84% and 80% at the three levels of granularity.[4] PDTB has served as the main

[3]This information is obtained via private communication.

[4]Prasad et al. (2008, p. 4): "The PDTB corpus was sense annotated by two annotators, with inter-annotator agreement computed for the three tag- ging levels. At class level, we noted disagreement when the two annotators picked a subtype, type or class tag of differ- ent classes. At type level, we noted disagreement

data source for training discourse relation classification systems since it was published. These studies will be reviewed in the next section while I introduce linguistic features of the discoure relations. A review of the application-oriented studies on this data has been done by see Webber and Joshi (2012). In my study the corpus is used for the first time to validate a set of psycholinguistic hypotheses about the comprehension and production of discourse connectives.

2.4 Markers of discourse relations

This section is dedicated to the review of the computational attempts to identify linguistic markers of the relations including expressions that are traditionally classified as discourse connectives (Schiffrin, 1988; Fraser, 1990; Blakemore, 2002), as well as other signals such as syntactic, semantic and clausal features of the relational arguments that appear in more recent approaches. Machine learning attempts to discourse parsing makes use of these features to identify discourse relations in text. This category of previous work is reviewed to prepare us for answering the two main questions of the thesis, i.e., the questions involved with specificity of the inferences when readers encounters a discourse cue (studied in Chapter 4) and the optionality of a discourse connective when speakers encode their messages in a multi-sentence utterance (studied in Chapter 5).

2.4.1 Discourse connectives

The basic function of this category of words and phrases is establishing connections between discourse segments and marking specific relations in text and utterances. What distinguishes them from other signals of discourse relations is that discourse connectives do not contribute to the propositional meaning of the individual discourse segments that they connect, they only affect the interpretation of the composition.

Table 2.1 demonstrates different syntactic categories of discourse markers. The last row of the table includes examples of phrases that appear with a relatively lower frequency than that of *syntactically admitted* categories of words, as put forth by Prasad et al. (2010).

when the annotators picked different types of the same class, e.g., Contrast vs Concession. Cases when one anno- tator picked a class level tag, e.g., COMPARISON, and the other picked a type level tag of the same class, e.g., Contrast, did not count as disagreement. At the sub- type level, disagreement was noted when the two annotators picked different subtypes, e.g., expectation vs. contra- expectation. Higher level disagreement was counted as disagreement at all the levels below. ”

Marker type	Examples
Conjunctions	*because, while, and, but, either..or, if, after, although*
Prepositional phrases	*despite, as a result, on the one hand..on the other hand*
Adverbials	*then, however, instead, yet, subsequently, too, eventually*
Cue phrases	*what's more, that is why, it was due mainly to*

Table 2.1: Discourse connective syntactic categories and examples

While discourse connectives are very informative about discourse relations, some challenge is involved with their identification:

Different readings: some words and phrases can function either as a discourse connective or as an intra-sentential argument depending on the context. Stede (2011) exemplifies *for* and *as long as* which would not convey any discourse-level information when used with non-clausal arguments, as in (10), while they can be very good discourse markers in some other context, e.g., (10-a). Therefore, classification of these phrases into sentential vs. discourse elements is itself a separate task and this is where the syntactic patterns need to be employed (Litman, 1996; Marcu, 1997).

(10) a. The lyrics I wrote **for** this song is **as long as** the previous one. – *sentential reading*

b. **As long as** (you love each other), ((nothing really matters), **for** (you can talk about the problems.)) – *discourse connective reading*

Semantic ambiguity: even when a word or phrase functions as a discourse connective, it can be ambiguous in three different ways. Firstly, some connectives have multiple discourse level readings: *since* can function as a temporal operator or a causal one (11).

(11) a. I never saw him again since we met in Berlin. – *temporal*

b. We never met again since we both knew it would hurt. – *causal*

It also happens that a connective holds information about two or more different relation senses. Typical examples are temporal connectives such as *when* and *while*, respectively, exhibiting senses of conditionality and contrast besides marking a synchronous temporal relation (12).

(12) a. Ask about the solution when you have already tried to solve it yourself. – *synchrony* and *condition*

b. It bothers me that he's wandering in facebook while I'm spending my whole day on finding a good apartment. – *synchrony* and *contrast*

The third type of semantic ambiguity arises from the generality of a connective that is applicable to a wide range of discourse relations (13). This type of ambiguity depends also on the adapted relation taxonomy. For example, if we only want to classify relations into positive and negative polarity relations, i.e., whether the two sentences are talking about congruent or incongruent events, then both following examples of *but* fit into the second category. In a finer-grained classification they should rather be categorized as two different relations.

(13) a. John loves Mary but she pretends to ignore him. – *concession*

b. Julia is tall but Jenny isn't. – *contrast*

All these types of ambiguities show that discourse connectives and discourse relations are different linguistic phenomena. We cannot reduce one to the other: a given connective type can be used in a variety of relations and, as we will see in the following sections, relations can be expressed by the help of other linguistic devices as well.

2.4.2 Lexico-semantic features

The propositional meaning of individual arguments of a relation is built upon the meaning of its words and phrases. Therefore, it would not be surprising if co-occurrence of a pair of words in two neighboring sentences constructs a discourse-level relation between the two propositions. Word-pairs (where each word comes from one of the involved arguments) are the most common representation of lexical features used for automatic identification of discourse relations. See how the relations between lexical items in (14) trigger relations between sentences that I have collected from natural text on the web.

(14) a. One of the **best features** of the iPhone 5 is its Wi-Fi. However, some users have argued that the Wi-Fi is **awful**. – *contrast*

b. Successful people never care about **others**, they are more cleared about **their** goals and **their own** uniqueness. – *alternative*
c. There has been an increase in the level of **debt** and often **uneconomic projects** has been financed. These have led to increasing **concern**. – *result*

Lexical features of this type are more ambiguous indicators of a relation compared to discourse connectives. For example, by keeping a similar set of words in both arguments of the exemplified relations in (14) and applying only slight changes different discourse relations can be obtained (15).

(15) a. Successful people care about **others** and help them. They are focused on **their own** benefits as well. – *expansion*
b. There has been a decrease in the level of **debt** and funding of the **uneconomic projects** has been cut. These are done to deal with the current **concern**. – *reason*

Nevertheless, word-pairs have proven to be helpful features in attempts for identification of relation sense both in the absence of discourse markers (Wellner et al., 2006; Pitler et al., 2009; Lin et al., 2009; Zhou et al., 2010; Park and Cardie, 2012; Wang et al., 2012) and in collaboration with them for classification of explicit relations (Marcu and Echihabi, 2002; Lin et al., 2010; Wang et al., 2010; Hernault et al., 2011; Versley, 2011a, 2013; McKeown and Biran, 2013; Rutherford and Xue, 2014).

Some studies emphasize on the role of content words for constructing lexical features with the intuition behind the above mentioned examples (Marcu and Echihabi, 2002), but it has been repeatedly shown that function words in the arguments of a relation also convey some information about the relation sense. Specifically, throughout comparative experiments of Blair-Goldensohn et al. (2007) and Sporleder (2008), performance of the models are superior in the condition that function words are included for pair construction. Pitler et al. (2009) show that even word-pairs such as *the-it* and *a-the* provide some information about contrast and contingency relations, respectively. McKeown and Biran (2013) present accuracy of classification separately for different types of relations and find that function words are in particular helpful for the detection of temporal relations. They attribute this result to the effect of tense markers such as *will* and *was*. This is in line with the findings of Lapata and Lascarides (2004) ons within-sentence temporal relations. While in their

study verbs are found to be one of the best features, no useful information is obtained from nouns and adjectives for prediction of the temporal ordering between clauses.

Lexical devices of discourse relations have been tried also in some other configurations besides cross-argument word-pairs, such as:

- n-gram pairs extracted from both arguments (Sporleder, 2008; Pitler et al., 2009; Versley, 2013),
- some words from the beginning and the end of arguments that might behave as cue phrases (Wellner et al., 2006; Pitler et al., 2009; Zhou et al., 2010; Park and Cardie, 2012; Wang et al., 2012),
- intra-argument n-grams again to capture phrasal features (Zhou et al., 2010; Wang et al., 2012).

Lexical features are the approximation of more abstract semantics that enables the reader to infer relations between sentences. The above examples should have made it clear that the worldknowledge is the basis for inferring relations. Capturing this type of information usually demands employing extra semantic resources such as lexicons and ontologies, or collecting unstructured semantics from large body of unlabeled but similar text (the distributional approach). Here, I review the attempts to employ semantic resources in automatic relation identification to mimic human inference based on worldknowledge.

Wellner et al. (2006) uses the Brandeis Semantic Ontology (Pustejovsky et al., 2006) to calculate semantic similarity between pairs of words in the arguments of a relation according to the distance (lexical path) between the corresponding entries in the ontology. They argue that this information should be helpful in identification of causal relations, e.g., when *crash* and *injure* occur in vicinity. The results indicate that this information is only as helpful as event head word pairs (which in most cases means the main verbs), leave alone that ontologies are expensive resources and usually cover semantic information only from a specific domain.

Sporleder (2008) compares different back-off models, i.e., lemmas, stems, sense disambiguated lemmas, and hypernyms from WordNet which is a valuable semantic resource to find relations between words. She finds that the morphological generalization on content words, namely, stemming and lemmatization, can be more beneficial than semantic back-off which makes use of external resources. Specifically in her experiments, WordNet hypernym back-off does not outperform the morphological generalization. The author attributes this result to the errors involved with the word sense disambiguation

system, which cannot be avoided. Feng and Hirst (2012) use different WordNet-based similarity measures to compute an average similarity score for the arguments of a relation based on comprising word-pairs and also considers VerbNet (Fillmore et al., 2003) to specify whether a class ID appears in either of the arguments or in both.

As a semantic back-off model for generalization over verb types, *Levin* verb classes (Levin, 1993) and *Inquirer* tags (Stone et al., 1966) have also been used in discourse relation identification (Pitler et al., 2009; Zhou et al., 2010; Park and Cardie, 2012). Other hand-crafted lexico-semantic categorizations such as words related to money, numbers and percentages are employed in (Pitler et al., 2009). They are found to contribute some information but are very genre-specific.

One issue attached with the use of external lexicons such as Levin classification, Inquirer tags, WordNet, etc. is that the semantic information obtained from these resources might not pertain to the type of world knowledge required for understanding of the text under study. This goes back to the necessity for word sense disambiguation. Many words have different senses and the information about a wrong sense, i.e., too general, too specific or totally unrelated to the context would not help at all for modeling the propositional meaning of a sentence.

Recently more effort is shaped around distributional semantic models which can be learned on desired data and have built-in strategies to deal with sense disambiguation. A distributionally obtained vector-space also offers homogeneous representation for semantic composition at different levels of granularity, e.g., the meaning of a clause can be compared against the meaning of a multi-sentence text span. In order to incorporate distributional semantic features, the typical approach is to train a model on large text corpora — possibly of the same genre to the target text — and then employ learned vectors as a substitute for word meaning. Rutherford and Xue (2014) employ a method of clustering over a semantic vector space to tackle the sparsity problem for word-pair features. Therefore, instead of taking the actual word-pairs as a feature for classification of discourse relations each word is mapped to its *Brown cluster* (Brown et al., 1992). Clusters are constructed via distributional processing of a news corpus of 63 million words and they generalize over various types of lexical items (function/content words, named entities, and even signs and numbers). This approach for the first time shows the potential superiority of semantic features compared against raw lexical representations, as well as syntactic features in delivering information about discourse relations.

2.4.3 Clause-level functional features

Lexico-semantic approaches are easy and intuitive choices for discovering discourse relations, but they only provide a very abstract understanding of discourse level inference. In order to approximate the meaning of a clause one should also consider the elements that are directly related to the propositional semantics. Sentence polarity, tense and modality are the three important features usually extracted from each argument of a discourse relation to represent shifts in a discourse, and in turn, to identify the discourse relations (Pitler et al., 2009; Zhou et al., 2010; Versley, 2011a; Park and Cardie, 2012; Wang et al., 2012; Versley, 2013).

Wellner et al. (2006) look at pairs of events appearing in the two arguments of a relation, their attributes, and the temporal relation between them extracted by the TARSQI system which is a TimeML-type event mention annotator. Event-pairs make a more tangible device to establish discourse relations compared with verbs, because events connect directly to the real-world phenomena and propositional meaning of a sentence. Along with every event mention extracted from the text, TARSQI provides some coarse-grained classification of the event types: occurrences (walk), reporting (tell), perception (observe), etc. Wellner et al. show that considering this information along with the tense, modality, temporal and subordinating links between event mentions of the two arguments (while this one happens rarely for relations other that *attribution*, e.g., *Mary saw that John left the party*) has a positive effect on the relation identification accuracy.

Wang et al. (2010) investigates the effect of temporal ordering of events separately in sense recognition of implicit and explicit relations. While these features are useful for both implicit and explicit relation recognition, implicit relations benefit more from the temporal information encoded in the events. Unfortunately, none of the two mentioned studies investigate the effect of event features on discourse relations of certain senses separately.

To understand how event related information can be connected to the individual words of the involved sentence, consider the following example. In (16-a) once it is detected that *looking for something* and *have lost something* are both related to the same entity, i.e., *the girl*, the typical causal relation between the generic event types can be generalized to this instance and the causal relation between the two sentence can be inferred. The same applies to the contrast hidden in the *rise and fall* pair in (16-b) which has an important share in producing an alternative or concession relation between the explained situations: something was expected about the stock and something else happened to it.

(16) a. The girl was looking for something on the pavement. She had apparently lost her ring. — pragmatic cause

b. The management was expecting a rise in the stock and it rather fell over 40%. — concession

Currently, event extraction and argument alignment are also very young trends in NLP. Relations like the one we exemplified are not captured perfectly if we look at content (e.g., word-pairs) and form features (e.g., tense) separately. The following section takes us one step closer to the notion of structured semantics across sentences which reflect a discourse-level connectivity.

2.4.4 Referential devices

Moving from the propositional contents of the individual relational arguments, the next element in the sentences which intuitively should play an important discourse role would be coreference chains. In fact, entity relations have been identified as a complementary aspect of coherence to semantic discourse relations — for example, see *Centering* theory by Grosz et al. (1995).

Prasad et al. (2008) report that about a quarter of sentences in the PDTB section of the Wall Street Journal establish some sort of entity links with their local context, including cases where no coherence relation applies. Contrary to the orthogonal perspective towards the discourse effect of entity relations with that of coherence relations, Louis et al. (2010) observe interesting differences among discourse relation senses with respect to the way entities are referred in their two arguments. A variety of different encodings of entity features are examined in this study and some meaningful correlations between entity relations and discourse relations have been observed. For example, Louis et al. (2010) find that temporal relations have a lot of coreferent entity mentions in their arguments. Causal relations tend to have larger number of pronouns in their second arguments, and in comparison relations, the two arguments contain entities with similar but not an identical referents (17).

(17) **Longer maturities** are thought to indicate declining interest rates. **Shorter maturities** are considered as a sign of rising rates because portfolio managers can capture higher rates sooner.

The authors combine manual annotation of relations from Penn Discourse Treebank (Prasad et al., 2008) with that of entities from Ontonotes (Pradhan et al., 2007) to find an upper bound for the contribution of the entity features in identification of implicit relations. Their findings indicate that despite of the information encoded in the entity patterns about the relation sense, these devices are only as helpful as simple word-pair features, and in fact, would not increase the accuracy of classification when considered together with lexical features. Louis et al. suggest as a future direction that bridging anaphora (Asher and Lascarides, 1998) might have a stronger signaling effect for specific discourse relations — see (18).

(18) a. We were cleaning **the kitchen** on the weekend. Maria found a ring under **the cabinets**. – *entity-entity association*

b. The SFB project proposal is **accepted**. We got the **notification** on Friday. – *event-entity association*

Currently no reliable automatic strategy for detection of bridging in text exists. In theoretical research on coreference, not much agreement has been obtained regarding the definition of identity and implicit semantic associations between discourse entities (Hovy et al., 2013). Relations of different types between words such as synonymy, hypernymy, meronymy and etc. from semantic resources like WordNet (Fellbaum, 1999) might serve as an approximation of anaphora in its various shapes when pairs of words from relational arguments are examined. Lexico-semantic links between words in WordNet or distributional semantic similarity measures also provide a framework to consider semantic relations between words of different syntactic categories (e.g., verbs and nouns) for approximation of event/entity coreference (Asr et al., 2014).

An experimental study by Kehler et al. (2008) provides more evidence for relational coherence and coreference interaction. In a sentence completion task they find that the type of discourse relation between two sentences affects the processing of referring expressions. In particular, resolution of a pronoun to one of the previously mentioned entities is performed in different ways depending on the discourse relation. Object pronouns (19-a) & (19-b) tend to be interpreted as referring to the object antecedents in what they call a parallel relation but to the subjects in a result relation. The effect is altered for subject pronouns (19-c) & (19-d). This observation cannot be explained by previous accounts of pronoun resolution and is indicative of the connection between the semantic discourse relation and the chain of entities being mentioned in two clauses.

(19) Samuel threatened Justin with a knife, and...

a. Erin blindfolded him (with a scarf). (parallel relation)
b. Erin stopped him (with pepper spray). (result relation)
c. he blindfolded Erin (with a scarf). (parallel relation)
d. he alerted security (with a shout). (result relation)

2.4.5 Subjectivity and polarity

Substantial proportions of sentences in news, commentary and product reviews (all common genres of text in NLP) consist of opinions about incidents, people, things and ideas. Sentences in subjective text often comprise positive or negative sentiments and it affects the type of discourse relation that a sentence establishes with its context.

The MPQA lexicon (Wilson et al., 2005) of negative/positive/neutral polarity words has served as a resource for automatic identification of discourse relations (Pitler et al., 2009; Zhou et al., 2010; Wang et al., 2012; Park and Cardie, 2012). In all of these works, polarity of a sentence is obtained by simply adding up the polarity of every sentiment word (and countering the polarity in case negation is used). While polarity features boost system performance against the random baseline, they tend to be among the least informative devices according to previous findings.

Pitler et al. (2009) closely examined their training set of implicit comparison relations where they expected opposite polarity between the two arguments of the relation, and surprisingly found that this occurred only in 30% of samples. On the contrary, opposite polarity arguments were found in 52% of the causal relations. This all might be due to imprecise calculation of polarity at the level of the sentences.

2.4.6 Syntactic patterns

Syntactic information has proven to be very helpful in classification of both implicit and explicit relations. Production rules, for example, have repeatedly been found to stand on top of other features in ablative analyses (Zhou et al., 2010; Park and Cardie, 2012). Also, dependency parse rules and lexico-syntactic context around the mid point of relational arguments have been used for finer-grained classification (Lin et al., 2009; Hernault et al., 2011). Two examples of sub-trees frequently occurring in temporal and causal relations similar to the following are depicted in Fig. 2.4.

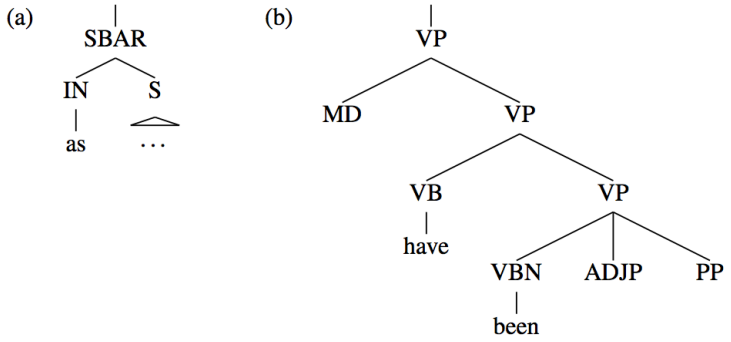


Figure 2.4: Two frequent syntactic structures appearing in Arg2 of temporal relations (a) and Arg1 of causal relations (b) (Lin et al., 2009)

(20) a. But the RTC also requires "working" capital to maintain the bad assets of thrifts that are sold. [subsequently] That debt would be paid off as the assets are sold. — *temporal* with implicit connective inserted

b. It would have been too late to think about on Friday. [so] We had to think about it ahead of time. — *causal* with implicit connective inserted

According to the accuracy obtained by applying individual feature sets by Lin et al. (2009), contribution of production rules (both internal and lexicalized ones) is superior to that of simple word-pair features (extracted directly from training data). Dependency information turns out to be less informative but authors attribute this to the errors of the dependency parser employed for obtaining these rules, whereas constituent parse trees have been obtained from the gold-standard annotations.

Wang et al. (2010) argues that syntactic information obtained from flat paths and second level production rules, as employed in the above mentioned works, can only capture part of the relational structure. They propose a kernel-based model to compare the parse trees of the relational arguments directly, thereby, capturing more syntactic information. Their system, which benefits from a set of baseline features in addition to the structured syntactic information, outperforms an equivalent considering the baseline features plus production rules by 7% accuracy. Their experimental findings on different sets of relations indicate that more sophisticated syntactic features enhance relation sense disambiguation specially when the two arguments of a relation are far apart in the text.

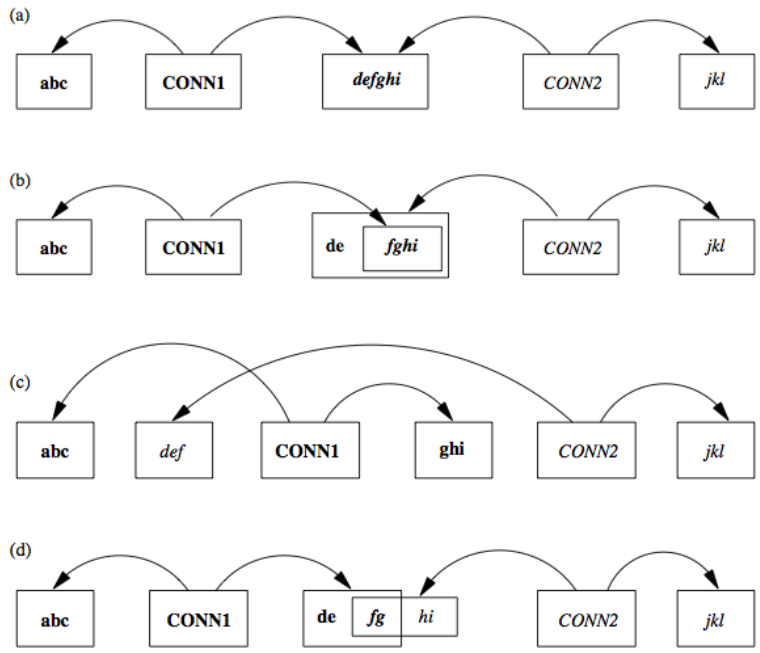


Figure 2.5: Different interconnectivities between arguments of neighboring discourse relations in text (Lee et al., 2006)

2.4.7 Context and neighboring relations

Like other linguistic phenomena, discourse relations in text interact with their context. One of the key ideas behind discourse parsing in the sense developed by the pioneers of the RST is that all pieces of a text are connected via relations, i.e., a meta structure on top of binary discourse relations is necessary for coherence of the text as a whole. However, the general methodology of RST in considering only the immediate discourse segments to shape new relations is a bit too strict and does not always apply to natural text. Lee et al. (2006) show that relational arguments shape a variety of inter-connectivities that are depicted in Fig. 2.5. Some relations have overlapping arguments, while some are totally separate from the text spans of their neighboring relations. Such adjacency features along with the sense of neighboring relations have been identified as influential information in determining the sense of a relation in its context (Pitler et al., 2008; Lin et al., 2009).

According to Pitler et al. (2008), certain pairs of discourse relations tend to occur together. For example, explicit comparison relations are usually followed by implicit contingency relations, whereas expansion relations occur in bunches (see Fig. 2.6). For the classification of implicit relations, Lin et al. (2009) use the connectives appearing in the preceding and/or following sentences of the relation as one feature in addition to the

First Relation	Second Relation	χ^2	p-value
Other	Other	66.2	$< .000001$
Other	I. Expansion	30.5	$< .000001$
I. Expansion	Other	20.2	.000007
E. Comparison	**I. Contingency**	**20.1**	**.000007**
E. Comparison	E. Comparison	17.4	.000030
I. Contingency	Other	14.8	.000120
Other	I. Contingency	13.6	.000228
E. Comparison	Other	13.3	.000262
E. Comparison	**I. Expansion**	**9.91**	**.001614**
I. Temporal	**E. Temporal**	**9.42**	**.002141**
I. Contingency	**E. Contingency**	**9.29**	**.002302**
I. Expansion	I. Expansion	9.09	.002569
Other	E. Expansion	6.37	.011567
I. Expansion	**E. Expansion**	**6.34**	**.011783**
I. Temporal	I. Expansion	5.52	.018784
E. Expansion	**I. Expansion**	**5.50**	**.019050**
I. Comparison	I. Expansion	5.45	.0195
I. Contingency	**E. Comparison**	**4.95**	**.0260**
E. Temporal	E. Contingency	4.24	.039571
E. Contingency	Other	4.15	.041728
I. Expansion	I. Contingency	3.93	.047475

Figure 2.6: Frequent adjacent relation pairs in Penn Discourse Treebank extracted byPitler et al. (2009)

type of argument inter-connectivities. In comparison to syntactic and simple word-pair features, contextual dependencies and markers of neighboring relations contribute less information about the relation sense. We also ran a preliminary experiment on Penn Discourse Treebank (Prasad et al., 2008) using the gold-standard senses of both implicit and explicit relations in the neighborhood composed with the argument dependency as more sophisticated contextual features (e.g., *both arguments embedded in a cause relation*) but did not obtain a better classification accuracy. Nevertheless, contextual features that (Feng and Hirst, 2012) extract from global trees in RST-DT (Carlson et al., 2003) corpus significantly boost the accuracy of their system in relation sense classification.

2.4.8 Modality-specific features

Contextual and orthographic features such as the neighboring punctuations, distance between the two arguments of the relation and position with respect to the paragraph boundaries constitute another class of relational cues that can only be found in text.

These features are employed in previous work and each contributes to the relation sense classification to some extent (Wellner et al., 2006; Pitler et al., 2009; Wang et al., 2010; Versley, 2011a). If relational inference is studied in other modalities such as spoken dialog, other features can be indicative of the relation type between discourse segments. Murray et al. (2006), for example, extract 75 prosodic features from speech data and use them for relation classification.

2.5 Summary

Discovery of relations among discourse segments in a text is a complex process which requires understanding the propositional content of the sentences and composing them to infer new information. This chapter provided a background overview of the work on definition of discourse relations, their place in the theories of discourse and finally the automatic approaches to identify discourse relations in text by the help of linguistic features. Despite of decades of research on this topic, we still don't have access to accurate discourse parsers and not even to manually annotated corpora with high annotation agreement scores. The most popular resource in the community is the PDTB corpus that we introduced in this chapter. Automatic discourse parsers developed on this data were examined in the CoNLL shared task this year (Xue et al., 2015).[5] The best reported accuracy of relation sense identification even without error propagation form other tasks (discourse connective and segment detection) is very low: 65.11% overall, 90.00% for explicit and 42.72% for implicit relations. Not only implicit relations (that contain no overt discourse marker) are difficult to identify by relying on the introduced linguistic features, but also explicit relations are sometimes difficult to be labeled with fine-grained relation senses due to the semantic ambiguities associated with the discourse markers. These results suggest that more fundamental research needs to be done on discourse relations and their markers.

In addition to proposing a new theoretical framework for studying discourse-level communication, the experiments we conduct on the PDTB corpus in this thesis elaborates on the differences between the contribution of discourse markers of various types, and the inherent properties of the implicit and explicit relations. This helps understanding the problem of automatic discourse parting in more depth, and hopefully would count as initial steps toward building computational models of human discourse comprehension and production in the future.

[5] http://www.cs.brandeis.edu/ clp/conll15st

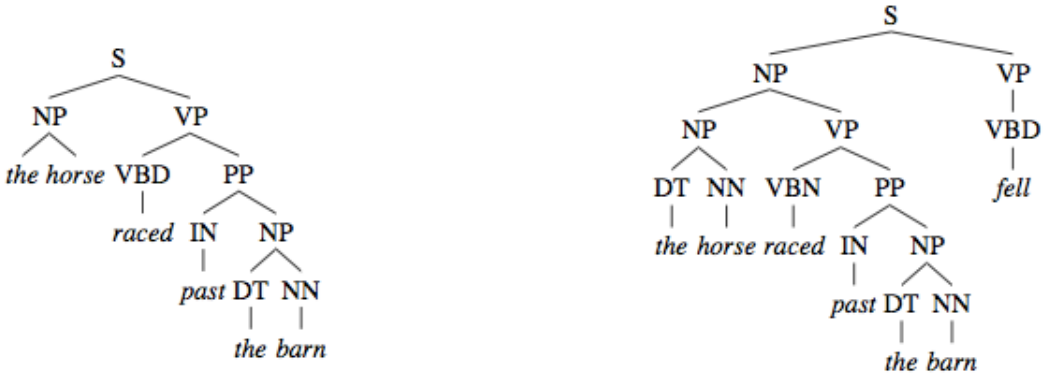
Chapter 3

A new framework for studying discourse relations

This chapter motivates an information theoretic study of discourse relations by reviewing previous work on other levels of human sentence processing. I discuss how employing ideas from information theory lets language researchers explain some phenomena in human communication that do not have a classic linguistic explanation. This approach enables us to make quantified predictions regarding the effect of discourse connectives on comprehension and production of multi-sentence text (that will be examined in the Chapters 4 and 5, respectively). An overall analysis of the Penn Discourse Treebank is conducted to show that discourse connectives vary a lot in terms of the amount and type of relational information they encode in text. Some connectives only appear in very specific relations, whereas some are widely used in a large set of relations. The main questions of the thesis regarding processing of ambiguous discourse connectives and production of implicit vs. explicit discourse relations will become more concrete after this analysis.

3.1 Information theoretic approach to communication

Besides many other factors, production and comprehension of language signals are governed by the principles of communication. For successful communication of a message, the speaker needs to choose the right form that will be interpreted by the listener the way it is intended. Comprehension on the other side of the channel also depends on the amount of uncertainty involved in decoding of the message's form into its meaning. Linguists have adopted these concepts from information theory, an applied mathematics framework developed by Shannon in 1948. The original theory aimed at solving problems in signal processing, cryptography, and data compression, but later broadened its application to many other disciplines including natural language processing and psycholinguistics. The primary motivation of information theory is the analysis of communication over a channel (such as a telegraph or Ethernet line) that can transfer certain amount of information per time unit or symbol. In natural language communication, speakers often have different options for encoding a message. The equivalency of meanings and diversity of forms



(a) Main verb reading

(b) Relative clause reading

Figure 3.1: Two readings of a garden path sentence from Hale (2001)

gives the speaker a chance to select the best form by considering other constraints that might affect interpretation of the message. For example, while talking in a noisy cocktail party people might speak not only with higher volume but also in slower pace to make sure that they will be heard correctly. In information theoretic terms, this means that less information is delivered in each time frame as the chance of misinterpretation goes higher, i.e., the channel capacity becomes smaller. Researchers in psycholinguistics have proposed different ways to study human communication from this perspective. The commonality of these studies is that they explain choices of speakers in the production side and listeners' comprehension difficulty based on the information content of the linguistic units that are put together to form a message. This is one way to study language in use (performance) when it can no more be constrained by the rules of grammar (competence).

3.1.1 Comprehension mechanism

One category of studies focuses on the comprehension aspects, trying to explain the difficulty involved with the processing of certain sentences based on the word-by-word delivery of information that occurs during incremental perception of these sentences. Hale (2001) proposes that the processing difficulty at a given word w_i in a sentence is proportional to the information that w_i adds to its preceding context and that this can be

measured by the *surprisal* of w_i :

$$\begin{aligned} S(w_i) &= -\log p(w_i|w_{1\ldots i-1}) \\ &= -\log \frac{p(w_{1\ldots i})}{p(w_{1\ldots i-1})} \\ &= -\log p(w_{1\ldots i}) + \log p(w_{1\ldots i-1}) \end{aligned} \tag{3.1}$$

where $w_{1\ldots i-1}$ indicates the string of words preceding w_i. The more likely a word is to appear in a context the less information it conveys to the reader. Therefore, according to the surprisal theory, a word would be difficult to process if it is not likely to appear in that particular context. The probability function in the formula can be computed based on purely lexical statistics (simple n-gram models) or more abstract representations that consider syntactic or semantic information. For example, the syntactic surprisal of a word can be obtained by introducing a new variable T indicative of the syntactic trees (see Levy, 2008; Demberg and Keller, 2008). Then the formula will be expanded to the following:

$$S_{syntactic}(w_i) = -\log \sum_{T \in Trees} P(T, w_1..w_i) + \log \sum_{T \in Trees} P(T, w_1..w_{i-1})$$

where the sum of the probabilities of all syntactic trees compatible with words $w_1..w_{i-1}$ appears in the first term and the sum of the probabilities of all trees additionally including word w_i appears in the second term. The probability of a syntactic tree given the lexical items can be obtained from a parser. Then modeling of the processing difficulty involved with specific syntactic structures becomes possible. A classic category of grammatical but difficult structures is that of *garden path* sentences (Frazier and Rayner, 1982) like the following:

(1) The horse raced past the barn fell.

Such sentences have been experimentally proven to be difficult to process during online reading because of a local ambiguity: a main verb reading of the *raced* is possible up until the reader reaches *barn* which is depicted in Figure 3.1a, while the correct reading involves a reduced relative clause parse as in Figure 3.1b. According to Hale (2001) the reduced

relative clause is seven times less frequent structure compared to the main verb structure in a sample of natural English text. The surprisal model easily captures the difficulty appearing at *fell*, given that its likelihood given the preceding syntactic context is so low. Surprisal has been used to explain the difficulty of processing complex structures within the boundary of the sentences such as object garden path sentences, relative clauses and long distance dependencies (Demberg-Winterfors, 2010). Semantic information at the sentence level has also been incorporated into some models of surprisal (Mitchell et al., 2010). Yet no model has been proposed to capture discourse level processing difficulty involved with comprehension of sentences like (2-a) when compared to more expected combinations like that in (2-b).

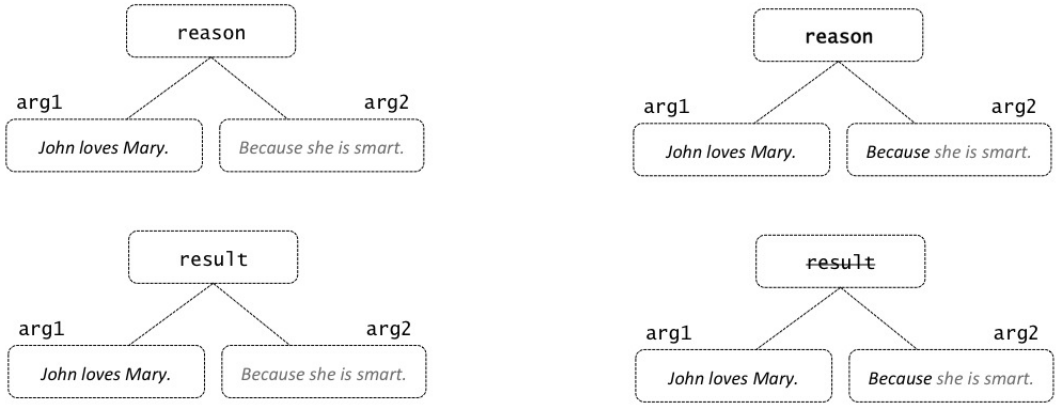
(2) a. ? Maria was terribly sick, therefore, she attended the course.
b. Maria was terribly sick, however, she attended the course.

If we aim at building high-coverage models of human sentence processing taking into account semantic dependencies beyond the boundary of individual sentences, we need a mechanism to deal with discourse relation information. The same way syntactic surprisal is constrained on production trees, we can constrain the word-by-word surprisal on the set of possible relational structures:

$$S_{relational}(w_i) = -\log \sum_{r \in Relations} P(r, w_1..w_i) + \log \sum_{r \in Relations} P(r, w_1..w_{i-1}) \quad (3.2)$$

where r is a minimal structure governing the entire text span of a relation and has a sense label. As with syntactic trees, we could say that a relational structure is either compatible or not with an utterance up to a given point w_i. However, this has to be determined with respect to the probabilities collected from a discourse annotated corpus. Some words — like very specific discourse connectives — distinguishably change the distribution, i.e., they increase the likelihood of some relations over the others. Figure 3.2 illustrates how a relation is removed from possible derivations when a highly informative connective is encountered. Other words, and in a more abstract way, other syntactic/semantic and clause level features of a sentence, can also be indicative of the discourse relation, thus changing the likelihoods in the above formula.

Building a generative discourse parser to provide us with word-to-word update of



(a) Before encountering the connective.

(b) After encountering the connective.

Figure 3.2: Likelihood of compatible discourse relations change as more words are encountered. Some words like *because* are highly informative and effective.

information on discourse structure is unfortunately not possible with the amount and quality of the available discourse annotated language data. In the future, when the annotation schema become more homogeneous and actual annotations obtain higher human agreement, and more data became available, then we might be able to develop computational models of human discourse processing based on surprisal and make specific predictions about processing difficulty of certain discourse structures the same way probabilistic parsers have been used for calculating surprisal at the level of syntax and making computational predictions about human sentence processing.

One objective in my thesis though, is to apply some simplifications to the above problem and use currently available data for specific predictions regarding the effect of connectives on comprehension of multi-sentence text. In order to do so, we use discourse annotated data for calculating the information content of a discourse connective. Before moving to that section, we have a look at the related work on production, too.

3.1.2 Production mechanism

One of the earliest accounts of language production is that of Zipf (1949). Zipf proposes that the distribution of words and structures in a language is a natural consequence of human tendency to apply the *principle of least effort* in communication. For example, more frequent words such as articles and prepositions tend to have shorter length than less frequent words. This idea shares some similarity with more recent theories of communication in pragmatics. In particular, Grice (1975) proposes a set of cooperation principles that are fulfilled in a successful communication, including maxim of quantity,

quality, relation and manner. The maxim of quantity, which is relevant here, states that:

> *Speakers should make their contribution as informative as is required for communicating their messages, and do not contribution more information than is required.*

While the principle of least effort simply accounts for efficiency in terms of the amount of effort put into the production of linguistic signals (thus using shorter forms for uttering more frequent concepts), the communication perspective explicitly considers ease of processing at the comprehender too. The Uniform Information Density theory (Levy and Jaeger, 2007) formulates this idea in a quantified manner by adopting the notion of communication through a limited-capacity channel (Shannon, 1948). This theory emphasizes that information has to be produced in a way that can be processed easily. A listener can process a certain amount of information at a time, thus a rational speaker should try to produce information in a rate close to this capacity. Therefore, the UID theory proposes that a rational production mechanism should work in the following way:

> *Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal. Where speakers have a choice between several variations to encode their message, they prefer the variant with more uniform information density.* — Jaeger (2010)

From a cognitive perspective this is a stronger argument about language processing compared with what theories like surprisal propose to be functioning in comprehension. If production data turns out to support the above hypothesis then it means not only that the comprehension process is probability-sensitive but also that the language producers take this into consideration (perhaps subconsciously)[1]. The UID principle entails that overly informative units should be broken into several units so they fit in the channel capacity and be processed by the listener (avoiding peaks in the information density). It also predicts that non-informative units such as optional markers where they are redundant should be dropped (to avoid troughs in the information density). Recall the definition of the information content of a unit (e.g., a word) in the surprisal theory. The information content is formulated with respect to the predictability of the word given the preceding context. Surprisal has been used for testing UID at different levels of language production. As an example, we look into the study of "that" omission by Jaeger (2010). Some verbs in English such as *think* and *say* frequently occur together with complement clauses, whereas others such as *confirm* are not highly selective of their object argument position. Using complement

[1] cf. Jaeger 2010, p 25: "the term 'choice' does not imply conscious decision making. It is simply used to refer to the existence of several different ways to encode the intended message into a linguistic utterance."

clauses would be grammatical in both cases with or without the complementizer, *that*:

(3) a. My boss thinks [that] I am absolutely crazy.
b. My boss confirmed [that] we were absolutely crazy.

The UID mechanism would suggest that the speaker's choice of dropping "that" in naturally occurring speech should correlate with the predictability of the continuation structure. In other words, speakers should tend to keep "that" in place where syntactic processing would otherwise be difficult due to unpredictability. On the other hand, speakers are hypothesized to avoid redundancy, thus drop "that", in a context where the expectation for a complement clause continuation is high. By looking into a corpus of spontaneous speech Jaeger (2010) finds that the proportion of "that" omission is strongly correlated with the predictability of the continuation in terms of how often a given verb phrase governs a complement clause structure. The word-by-word information delivery is depicted in Figure 3.3).

More evidence on the predictions of UID, i.e., that speakers choose among meaning-equivalent alternatives the ones that correspond to a more uniform rate of information transmission, has been provided by a range of recent experimental and corpus-based studies at the level of spoken word duration and articulation (Buz et al., 2014), morphology (Kurumada and Jaeger, 2013), syntax (Jaeger, 2010), lexical choices (Piantadosi et al., 2011; Mahowald et al., 2013), referring expressions (Tily and Piantadosi, 2009; Kravtchenko, 2014), and across levels, e.g., effect from syntax and semantics on spoken word durations (Demberg et al., 2012; Sayeed et al., 2015). Again, discourse relations are absent in the information theoretic studies of production. This is what I will examine by looking into the patterns of discourse connective use in natural text. Some general observations are made in the following section regarding the frequency of implicit relations in the PDTB and optionality of the discourse connectives. In Chapter 5, we investigate whether discourse connectives are used in ways consistent with predictability of the relations they mark. If optional markers in language are used to adjust the uniformity of information density, we expect that connectives should be absent when relations they mark are predictable and should be present when the information they deliver is essential for inferring a particular relations.

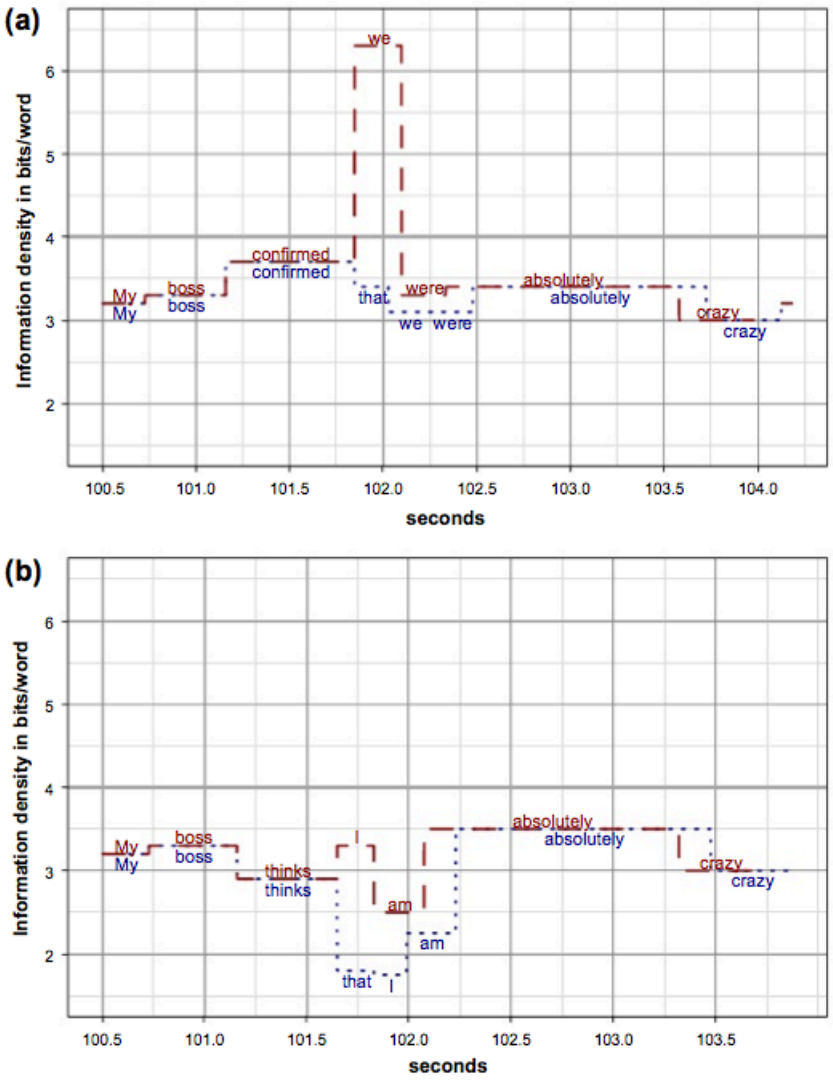


Figure 3.3: Information delivered by *that* in different contexts (Jaeger, 2013)

3.2 Communication via discourse connectives

When readers encounter a discourse connective in a text, they use it to infer a relation between the two discourse segments linked by the connective. Here it becomes important what type and amount of information is encoded in the connective. We propose that this information can be measured by looking at the distribution of a connective in natural text corpora annotated with discourse relations. We now examine the PDTB corpus introduced in Chapter 2, and calculate the information content of discourse connectives based on their co-occurrences with relations of different types. Throughout this section we make the following observations which give a direction to the later experiments on the effect of connectives' information content on comprehension and production:

1. Different connective types deliver different amounts of relational information.
2. Most discourse connectives are highly ambiguous, i.e., they can be used in a variety of discourse relations.
3. Most discourse relations can be signaled by more than one type of connective but some have their own unique markers.
4. Relation senses vary a lot in terms of how frequently they appear in text with or without a discourse connective.

These observations are in some ways dependent on the corpus we use for the analysis in terms of the type of text and its annotation schema, but this is an artifact of an empirical study on real data. The idea is to open a way for testing specific hypotheses about human communication at the discourse level. Later chapters connect these observations to previous theories and experimental psycholinguistic studies.

3.2.1 Measures of information

From an information theoretic perspective, markers of discourse relations remove uncertainty about the type of relation between two sentence. We take discourse connectives as a category of markers co-occurring with relations of different types. The mutual information between two discrete variables is indicative of the amount of uncertainty that one removes for inference of the other; Thus it can be used to capture the effect of a connective in its relational context. The mutual information between two variables c (connective type) and *r* (relation sense) is obtained by the following formula.

$$I(X;Y) = \sum_{c} p(c) \sum_{r} p(r|c) \log \frac{p(r|c)}{p(r)}$$

the inner sum is known as Kullback-Leibler divergence or relative entropy between the distribution of relations $p(r)$ independent of the connective c and the distribution of relations $p(r|c)$ after observing c. The relative entropy thus quantifies how much encountering the connective c changes the distribution of possible relations.

$$gain(c) = D_{KL}(p(r|c)||p(r))$$

Intuitively, connectives that have similar meanings should be distributed across relations of different types in similar ways, whereas two connectives with very different meanings should have different distributions. What ends up in the information gain formula is the amount of relational information encoded in a connective, not the relatedness of the connective to a particular relation sense. The latter can be calculated by the point-wise mutual information:

$$pmi(c, r) = \log \frac{p(c, r)}{p(c)p(r)}$$

I will use this measure a lot throughout the thesis [2]. But for now we focus on the entire uncertainty removed by a connective type from a system of relation sense categorization.

3.2.2 Levels of granularity

In the PDTB relations are organized in a hierarchical format which makes it interesting to see how much information is conveyed by each connective type at each of the three levels of granularity. I define the measure of *enhancement* to formalize this notion:

[2]The point-wise mutual information of a connective and a relation sense is a logarithm function of the likelihood of c to r. In a Bayesian inference model, the likelihood can be used to obtain the posterior probability of a hypothesis by multiplying it with the prior function:

$$p(r|c) = \frac{p(c|r)}{p(c)} * p(r) \propto likelihood * prior \tag{3.3}$$

In this context, if the listener of an utterance has a prior expectation regarding the type of relation it will have with a following sentence, encountering the connective updates the expectation, i.e., the posterior probability. We will get back to the Bayesian formula in Chapter 5.

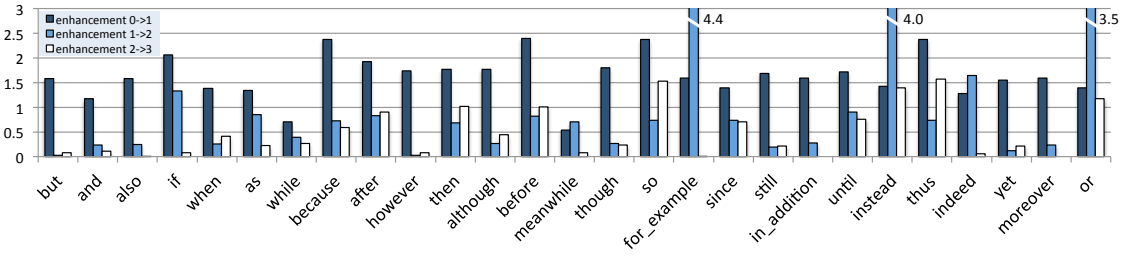


Figure 3.4: Information content of 27 most frequent connectives in the PDTB at three levels of relation sense classification — ordered left to right by connective frequency in the corpus.

$$enhancement_{x-y}(c) = gain_y(c) - gain_x(c)$$

The $enhancement_{x-y}(c)$ indicates the amount of information delivered by cue c for the classification of the instances into finer-grained relation sub-types after it has already been classified into a coarser relation. For example, $enh._{0-1}(because)$ describes how much information *because* provides for distinguishing the level-1 relations from one another, and $enh._{1-2}(because)$ is the additional information that this connective provides for distinguishing second level subcategories.

In order to examine the amount of relational information a connective delivers about relation senses in the PDTB, we extracted all *Explicit* relations from the corpus and measured information gain at three levels of granularity (see the hierarchy in Chapter 1). Figure 3.4 depicts the amount of enhancement for 27 frequent (> 100 occurrences) connectives in the corpus in three transitions, namely, from no categorization to the first level classification (`COMPARISON, CONTINGENCY, EXPANSION, TEMPORAL`), from first to the second level and from second to the third. Most of the connectives contribute most strongly at the coarsest level of classification, i.e., their $0 - 1$ enhancement is the highest. In particular, we find that some of the most frequent connectives such as *but*, *and*, and *also* only help distinguishing discourse relation meaning at the coarsest level of the PDTB relation hierarchy, but contribute little to distinguish among e.g. different subtypes of `COMPARISON` or `EXPANSION`. Frequent markers of `COMPARISON` relations *but*, *though*, *still* and *however* provide very few information about the second and third levels of the hierarchy. Another group of connectors, *for example*, *instead*, *indeed* and *or* contribute significantly more information in transition from the first to the second level. These are specific markers of some level-2 relation senses. Among them, *instead* and *or*, markers of `EXP.Alternative.conjunction` and

Relation pair	#R1 (total)	#R2 (total)	#Pair	χ^2
T.Synchrony–CON.Cause.reason	507 (1594)	353 (1488)	187	1.08E+00
T.Asynchronous.succession–CON.Cause.reason	189 (1101)	353 (1488)	159	2.43E+02 ***
E.Conjunction–CON.Cause.result	352 (5320)	162 (752)	140	2.22E+02 ***
T.Synchrony–EXP.Conjunction	507 (1594)	352 (5320)	123	5.43E+01 ***'
T.Synchrony–CON.Condition.general	507 (1594)	70 (362)	52	1.67E+01 ***
T.Synchrony–COM.Contrast.juxtaposition	507 (1594)	77 (1186)	45	1.97E+00
T.Asynchronous.precedence–E.Conjunction	66 (986)	352 (5320)	36	1.15E+01 ***
T.Synchrony–COM.Contrast	507 (1594)	37 (2380)	28	9.55E+00 ***
T.Synchrony–COM.Contrast.opposition	507 (1594)	28 (362)	21	6.78E+00 **

Table 3.1: Double-tagged relations in PDTB: frequency among double-tagged relations (and in the entire corpus)

`EXP.Alternative.chosen alternative` respectively, even help more for the deepest classification.

Temporal and causal connectives such as *before*, *after*, *so*, *then* ,*when* and *thus* have more contribution to the deepest classification level. This reflects the distinctions employed in the definition of the third level senses which has a direct correlation with the temporal ordering, i.e., forward vs. backward transition between the involved sentences. In other words, regardless of whatever high-level class of relation such markers fit in, the temporal information they hold make them beneficial for the 3rd level classification.

There are also a few connectives (*if*, *indeed*, *for example*) that convey a lot of information about the distinctions made at the first and second level of the hierarchy, but not about the third level. The reason for this is either that the third level distinction can only be made based on the propositional information in the arguments (this is the case for the sub-types of the `Conditional` relations marked by *if*), or that the connector usually marks a relation which does not have any third level sub-types (e.g., *for example* is a good marker of the `EXPANSION.Instantiation` relation which does not have any child in the PDTB hierarchy).

A sum over enhancements obtained in the three levels results in the total relative entropy of the distribution of discourse relations *a prior* vs. *posterior* to encountering the connective.

3.2.3 Ambiguous connectives

Three different types of ambiguity associated with connective types can be observed in this data:

1. A connector expressing different relations, where it is possible to say that one but not the other relation holds between the text spans, for example *since*.

2. A connector expressing a class of relations but being ambiguous with respect to the sub-classes of that relation, for example *but*, which always expresses a `COMPARISON` relationship but may express any sub-type of it, such as `Concession` and `Contrast`.

3. The ambiguity inherent in the relation between two text spans, where several relations can be identified to hold at the same time.

The first and second notion of ambiguity refer to what we so far have been talking about: we showed that some connectors can mark different types of relations belonging to one or several coarse categories. These connectives exhibit smaller enhancements (see Figure 3.4). The third type of ambiguity is also annotated in the PDTB. Relations which are ambiguous by nature are labeled with two senses on which the annotators agree[3].

Table 3.1 lists which two relation senses were most often annotated to hold at the same time in the PDTB, along with their frequencies. Sub-types of `Cause` and `TEMPORAL.Asynchronous` relations appear most often together. Also, `TEMPORAL.Synchrony` is a label that appears significantly more than expected among the multi-sensed instances, with an even higher frequency than that of `EXPANSION.Conjunction`, the most frequent label in the corpus. Such observations confirm the existence of the third type of ambiguity in discourse relations. Interestingly, these inherently ambiguous or multi-sensed relations also have their own specific markers, such as *meanwhile* which occurs in about 70% of its instances with two relation senses[4]. On the other hand, other well-known ambiguous connectors like *since* rarely mark inherently ambiguous relations, and most often can be identified as one specific relation sense by looking at the content of the arguments. The importance of the possibility to annotate a second sense and hence explicitly mark the inherently ambiguous relations has also been pointed out by Versley (2011b). In fact, a connective like *meanwhile* can be thought of as delivering information not only about the possible relation senses it can express, but also about the fact that two discourse relations hold simultaneously.

[3]These might be called double-tagged relations, but we prefer to call the connective used in these relations as ambiguous in the sense that it does not directly guide us to a single relation.

[4]This connective is mostly labeled with `TEMPORAL.Synchrony` and `EXPANSION.Conjunction`. Interestingly these two labels appear together significantly less frequently than expected (as marked in the table with ***') but when such a co-occurrence happened in the corpus it was with the connective *meanwhile*.

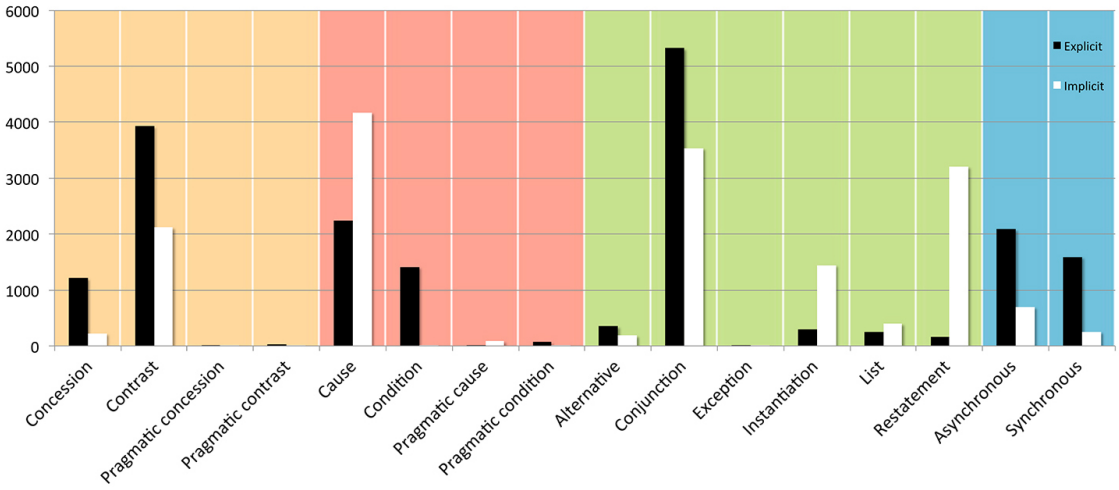


Figure 3.5: Distribution of level-2 relation senses in the PDTB across Explicit and Implicit occurrences.

In conclusion, there is quite a lot of ambiguity attached with most discourse connectives in terms of what relations they can mark. Of particular interest to us are connectives that appear in the same coarse-grained relations, such as markers of COMPARISON relations. We would like to examine how these connectives help inferring finer-grained relations and how their semantics should be defined given that they deliver some information about each sub-type of the coarse-grained class. This is one of the questions we might be able to answer by looking closely into the distributional differences between two connectives. Chapter 4 includes a case study on *but* and *although* to validate this claim.

It is important to emphasize that our analysis regarding the different senses of a connective and its information content is very much dependent on the PDTB hierarchy of relation senses. There is no ground truth in the design of this hierarchy. This means that a similar analysis on other corpora with different annotation schema should be theoretically acceptable, even though it might result in other observations regarding particular discourse connectives. What we nevertheless expect is that the findings of studies across corpora should not be contradictory, even if they are not the same. Furthermore, any computational modeling of the semantics of discourse connectives should be validated experimentally and/or discussed in the context of experimental studies, as we exemplify in the case study in the next chapter.

3.2.4 Absent connectives

In the PDTB, relations have been annotated also where no explicit discourse connective existed between two adjacent and related sentences. An interesting observation is that different relation senses from the hierarchy occur with very different frequencies across *Explicit* and *Implicit* annotated relations. Figure 3.5 shows the distribution of the relations (aggregated to their level-2 labels) in the corpus. Some relations never appear implicitly, such as the `Condition` relation. That can be explained according to the grammar involved with formulation of this type of relation: in order to indicate an statement being conditioned on another, one needs to use *if* in English. Other formulations are possible but very rare.

(4) a. If you had studied hard, you wouldn't fail the exam. — *Explicit (if)*
b. Study hard, and you wouldn't fail the exam. — *Explicit (and)*
c. Had you studied hard, you wouldn't fail the exam. — *Implicit (syntax)*

Among the most frequent senses, `Cause` and `Restatement` tend to appear without connectives, whereas `Contrast` and `Conjunction` are more frequent among the *Explicit* relations. Given that both formulations of these relation senses are possible (each of these senses occurs both with or without connectives in the corpus), the grammatical approach cannot explain why these relations are so different in terms of implicitation, because apparently, most relation types can occur both with and without a connective. In Chaper 5, we will investigate whether communication principles, and in particular the mechanism proposed by the Uniform Information Density theory play a role in the speaker's choice of using vs. dropping a discourse connective.

3.3 Summary

This section provided an introduction to the information theoretic approaches in psycholinguistics and motivated studying discourse-level communication in such a framework. We proposed that a discourse marker can be viewed as a linguistic symbol that delivers information about the semantic relations between discourse segments in text. By using information theoretic measures we developed a methodology for discovering differences between discourse connectives and explaining them in a quantified manner.

The analysis of discourse connectives in Penn Discourse Treebank revealed that some connective types carry more information about discourse relations than others. Some are more specific and only mark a unique relation, whereas others are distributed in a variety of relations. We identified three types of ambiguities: a connective can mark two relations at a time (*while* can mark temporal synchronous and contrast relations between events), or it can mark different relations depending on the context (*since* sometimes marks a temporal relations and sometimes a causal relation), or it can mark multiple sub-types of a general class of relations (*but* can be used in contrast or concessive relations which are sub-types of comparison or negative polarity relations). These observation shapes our first research question that needs to be answered experimentally: Does the distribution of a connective influence relational inferences? In other words, we want to see if there is a connection between the comprehension of a discourse connective and its probabilistic distribution in natural text.

We also found large differences among connectives regarding their presence in natural text. Some relations tend to appear with explicit connectives, and some tend to occur without their connectives. This suggests that in some contexts a connective (that can possibly be present) is omitted by the speaker, leading to the question why and in what context? Each of the following two chapters looks into one of the above questions and adds to our knowledge of discourse processing by providing a quantified view of the function of discourse connectives.

Chapter 4

Distribution of a connective affects its comprehension

Does the distribution of a connective across relations of different types tell us anything about how it is processed by language comprehenders? This is the question we try to answer empirically by looking into specific connective types in PDTB. We first see how they are distributed in the corpus, and then examine them in offline and online reading tasks. The following section provides a general overview of different approaches to discourse connectives as markers of discourse relations. Next, discourse comprehension processes are identified and explained in detail to prepare the reader for our experimental setup. Finally, two offline and one online reading experiments are conducted on *but* and *although*, which validate specific predictions about these two connectives regarding a corpus-based analysis. These experiments investigate the function of two multi-sense connectives with similar meanings in comprehension of short narrative text. Previous psycholinguistic experiments provide very limited and high-level understanding of how alternating the connectives could change the interpretations of a given story. Our experiments on *but* and *although* show that indeed the very fine-grained differences in the distribution of the two connectives show up in human sentence comprehension. This also means that a quantified model of discourse connectives based on natural production data is capable of predicting the comprehension behavior; thus the two processes are inter-related.

4.1 Previous approaches to connective meaning

As we saw in Section 3.2.3, not only connectives such as *since* and *while* need to be disambiguated by their context, but also a connectives like *but* which appears in different fine-grained relations of a major class can be viewed as ambiguous discourse markers. However, the common practice for defining the semantics of discourse connectives of this type has been to define a single meaning for them: find the most general function of the connective that is present in any context where it fits and describe it by examples. Following the terminology put forth by Fraser (1999), I call this a *core meaning* approach. Studies in this direction have proven that finding a universal meaning is not trivial for some

general connectives that appear in a variety of discourse relations.

In more recent pragmatic studies, connectives have been approached from the viewpoint of relevance theory (Wilson and Sperber, 2002). Deviating from a classic linguistic approach, this theory emerges from a cognitive perspective of language and looks into the processes taking place in the mind of a reader. Blakemore (2002) proposes a relevance-based account of discourse connective by saying that these elements lead the readers to an intended interpretation of a sentence that might not be made if the connective is absent. But again, the attempt to find the procedures of particular discourse connectives in Blakemore's work, as well as related studies (Iten, 2000; Hall, 2004), is very similar to the previous account that describes a certain meaning for a connective. Therefore, these approaches can also be classified as descriptive accounts.

The information theoretic perspective, on the other hand, views every discourse connective as a probabilistic multi-sense marker and brings attention to the frequency of each relation sense that co-occurs with it, rather than seeking for an underspecified meaning. This perspective entails that the more frequently a relation occurs with a connective, the more likely that particular relation is to be inferred if that connective is used in a new context (unless the context in not neutral in preferring one over the other interpretation). It also means that the fine-grained semantic distinctions across different usages of a connective should be considered as part of the meaning attached to the connective. This is an information theoretic approach looking into the probability of a linguistic symbol to be interpreted in different ways, rather than proposing a certain meaning or function for the symbol. Experiments in this chapter are designed to show that the predictions of our account about similarities and differences between connective types are more robust and accurate compared to those of the core meaning accounts. In particular, I would like to tease apart the effect of context, i.e., the two relational arguments and the connective on the relations that will be inferred. In the core meaning account, the varieties we see in the usage of a general connective such as *but* is attributed to the effect of context, saying that the connective only marks a coarse-grained relation between the two sentences. This means if the context is not highly selective (i.e., several interpretations can be made) and a general connective is used, then readers will make underspecified interpretations. The distributional account, on the other hand, predicts that different interpretations in such a context will be weighted by the likelihood of relations given the utilized connective. That is, the connective biases interpretations towards the more specific relation sense that co-occurs with it more frequently in natural text corpora. Experiments in this chapter are all designed to find these probabilistic effects on human discourse-level comprehension.

Previous psycholinguistic experiments have looked into the discourse effect of the connectives that are very different from one another (e.g., a causal connective like *so* vs. an adversative connective like *however*) to show how each category of connectives affects comprehension of the involved sentences (Murray, 1995; Millis and Just, 1994; Nieuwland and Van Berkum, 2006; Köhne and Demberg, 2013; Drenhaus et al., 2014). Of course, such differences are easily predictable by the distributional account: a connective like *so* has a totally different distribution of relations than that of *however*.

(1) Mary was feeling hungry. There was some cake and pizza leftover in the fridge.

a. She wanted something savory, so she had a piece of pizza/cake.
b. She wanted something savory, however she had a piece of pizza/cake.

The difficulty lies in saying how similar connectives affect sentence comprehension. For example, we do not have experimental evidence on the different cognitive procedures guided by *but* vs. *although*. In particular, previous experimental studies have not looked into such pairs of connectives to see whether one could replace the other in a neutral context without changing the interpretation or the processing difficulty. In order to prove the predictive power of the distributional account, we target very fine-grained differences between *but* and *although*. The two connectives occur almost always with `COMPARISON` relations in the PDTB corpus.[1] They differ in terms of their distribution across finer relation senses. Using each of these connectives in a context where both can perfectly fit enables us to see how different interpretations are obtained by only changing the connective and thus examine the distributional account against the core meaning account. In Example (2), the combination of the sentences is coherent in either condition; However, each connective generates a specific interpretation of the text, which in turn leads to a specific expectation of how the story should be continued.

(2) Mary was felling hungry.

a. She took some cake from the fridge, but she wanted something savory.
b. She took some cake from the fridge, although she wanted something savory.
... She had a piece of cake/pizza and went to bed earlier than usual.

[1] In PDTB *but* occurs about 3% of the time with the class `Expansion` too.

We evaluate the effect of *but* and *although* beyond the boundaries of the attached clausal arguments, i.e., on hidden inferences which lead to acceptance or rejection of the following context. The continuation sentence in the above short story informs the reader of what Marry eats in the end, and its acceptability is modulated by the connective used in the second sentence. The coherence judgment experiments in section 4.4 and 4.5 provides evidence for the distributional hypothesis. We find that an ambiguous connective biases interpretation towards the relation sense that it most frequently marks in PDTB, which results in lower acceptability of a continuation that is incompatible with that interpretation. In section 4.6, we design an eye-tracking experiment to look into the effect of discourse connectives on raising expectations during online processing of text for particular upcoming words, that is on *cake/pizza*. Results of the online experiment indicate that fine-grained differences between *but* and *although* are noticed in online reading but much less than in an offline task, i.e., the explicit coherence judgment.

4.2 Background on discourse comprehension processes

According to Kintsch (1988) *"discourse comprehension, from the viewpoint of a computational theory, involves constructing a representation of a discourse upon which various computations can be performed, the outcomes of which are commonly taken as evidence for comprehension. Thus, after comprehending a text, one might reasonably expect to be able to answer questions about it, recall or summarize it, verify statements about it, paraphrase it, and so on."*

I will try to sketch a clarifying overview of the previous psycholinguistic work on discourse connectives by categorizing them with respect to the comprehension processes they studied. In particular, the effect of connectives on three distinguished processes are reviewed: *integration* of new information with the preceding context, online *prediction* of the upcoming context, and most importantly, *inference* of new statements by putting together the content of the relational arguments in a specific way determined by the discourse connective. Integration is a bottom-up comprehension process, in which small pieces of an utterance are combined gradually with the context as they are perceived. On the other hand, prediction is the product of a top-down process, that is thinking ahead of the linguistic input. Integration and prediction have been studied vastly at the sentence-internal levels, i.e., for words and phrases within their lexico-semantic context (Wicha et al. (2004); DeLong et al. (2005); Van Berkum et al. (2005); Federmeier (2007); Otten and Van Berkum (2008); Van Petten and Luka (2012); Lau et al. (2013), among many

others). But here, we are talking about the semantics of a discourse cue and its effect on integration/prediction of a sentence, phrase or word in its discourse context. Experimental studies in this domain deal with a different class of cognitive processes, namely inferences. An inference only occurs at the level of discourse, requiring propositional units of language input, and is made based upon the comprehender's knowledge of events in real world. For example, one can infer a relation between the two sentences in (3-a) based on prototypical situations similar to what is happening for Harry. A discourse connective might trigger an interpretation that is either an emphasis or a deviation from how sentences might be interpreted in the absence of the connectives (see (3-b) and (3-c)).

(3) a. The boss was angry. Harry skipped the meeting.
b. The boss was angry, because Harry skipped the meeting. (Reason)
c. The boss was angry, so Harry skipped the meeting. (Result)

While all experimental studies on discourse connectives have something to do with the inferential processes, the majority only look into how easily the two sentences are integrated, or whether any part of the second sentence can be predicted regarding the connective. First, I will look into this category of previous work (Kintsch and Van Dijk, 1978; Just and Carpenter, 1980; Haberlandt, 1982; Kintsch, 1988; Millis and Just, 1994; Murray, 1995, 1997; Sanders and Noordman, 2000; Köhne and Demberg, 2013; Drenhaus et al., 2014; Rohde and Horton, 2014). Then I will focus on a slightly different question, that is what **additional statements** are inferred when two sentences are glued together by a particular connective. This is, the main question of the second set of studies I review in this section, after giving an introduction to linguistic inferences (Caron et al., 1988; Noordman and Vonk, 1992; Millis and Just, 1994; Cozijn et al., 2011; Wlotko and Federmeier, 2012).

4.2.1 Integration

Considering a connective with its two clausal arguments, a body of psycholinguistic experiments have explored the process of integrating new content (the second argument of the relation) with the old content (the first argument of the relation) by the help of connectives of different types. The measures to determine the effect of connectives on integration range from the reading time of the second argument in online reading to the posterior recall of the relation's content. A typical design to see how the connective influences these measures includes conditions with and without a suitable connective

(regarding the relation that is assumed to exist between the two sentences). The presence of a connective between neighboring sentences in a text reduces the reading time of the second clause, enhances memory recall of the story and accuracy in answering the comprehension questions in the studies of Haberlandt (1982); Millis and Just (1994); Sanders and Noordman (2000). The general hypothesis shaped based on these findings is that connectives facilitate integration of the second argument of the relation with the first argument.

Murray (1995) looks into different categories of connectives separately to further explore the equivocal perspective towards connectives of different types and refine the general hypothesis. He finds that differences in reading times and recall of contents are only significant when using adversative connectives for marking discontinuous relations, and not in cases where causal or additive connectives are used to signal continuous relations. In a similar vein, Murray (1997) examines reading of coherent sentence pairs with inappropriate connectives and finds a bigger disruption effect for adversative connectives than for the other two types of connectives. Putting together the findings of the two studies, Murray concludes that markers of discontinuous relations such as contrast and concession should have a more salient effect on discourse comprehension because they cancel the default expectation of continuity. Finding similar within-category effects and different between-category effects for markers of continuity vs. markers of discontinuity is a starting point for experimental studies to identify categories of connectives triggering similar procedures. This reminds us of the dimensions that are theoretically motivated and used for classification of discourse relations in the work of Sanders et al. (1992); Knott and Dale (1994) and Knott and Sanders (1998). The idea behind Sanders, Knott and colleage's proposal is that discourse relations and their markers should be classified according to the cognitive processes underlying a relation; for example, whether or not relating two sentences would need a causal inference (basic operation), whether the causal inference is semantic or pragmatic (source of coherence), and whether the causal relation is positive vs. negative (polarity). Only few experimental studies have been conducted to explore such fine-grained differences among connectives. Canestrelli et al. (2013) look into the Dutch connective *want* as a marker of subjective or pragmatic reasoning and *omdat* as a marker of objective and direct causality. They find that the difference in this dimension (i.e., source of coherence) indeed plays an important role in comprehension. The subjective connectives *want* induces a subjective representation (also know as a diagnostic relation), whereas the objective connective *omdat* triggers a direct semantic relation between the two events. If either of these connectives is used in place of the other, processing of the second sentence becomes more difficult (4).

(4) a. Hanneke was buiten adem, omdat ze vier trappen was afgerend om de post te halen. (Objective reason)
(Hanneke was out of breath, because she ran down four stairs to get the mail.)

b. Hanneke had haast, want ze was vier trappen afgerend om de post te halen. (Subjective reason)
(Hanneke was in a hurry, because she ran down four stairs to get the mail.)

This is evidence for the various procedures triggered by different markers of causality, and in turn, proves that categorization of relations and connectives based on the investigated dimension (source of coherence) is cognitively plausible. I will talk more about the basic dimensions of discourse relations in experiments of this chapter on comprehension and those of the next chapter on production processes.

4.2.2 Prediction

A general discussion among psycholinguists regarding integration of sentences within a larger context is involved with the question of incrementality. One view is that listeners initially compute the meaning of words and phrases with regard to local syntactic and semantic information before integrating them into the broader context, i.e., connecting a clause to the preceding sentences (Kintsch and Van Dijk, 1978; Kintsch, 1988; Just and Carpenter, 1980; Millis and Just, 1994). This hypothesis explains the wrap-up phenomenon (increased reading time at the end of sentences) observed in reading experiments. In particular, Millis and Just (1994) propose a *delayed integration* effect for the connectives according to which the content of the first clause is reactivated in memory when reading the second clause is finished and only then the content of the two are integrated. This hypothesis is based on observations from a self-paced reading task: readers processed the words appearing after a causal or adversative discourse connective quickly, compared to a no-connective condition, but then they slowed down at the end of the second sentence. Reading times were measured during a self-paced reading experiment where subjects strike a key to make the text appear on the screen word-by-word or in larger chunks. To investigate the time-course of processing discourse relations in a more natural setting, Traxler et al. (1997) conduct an eye-tracking experiment and collected reading time measures to analyze what parts of the text were more difficult to read. This time the stimuli consists of two conditions with objective reason vs. diagnostic relations both utilizing *because* as their connective.

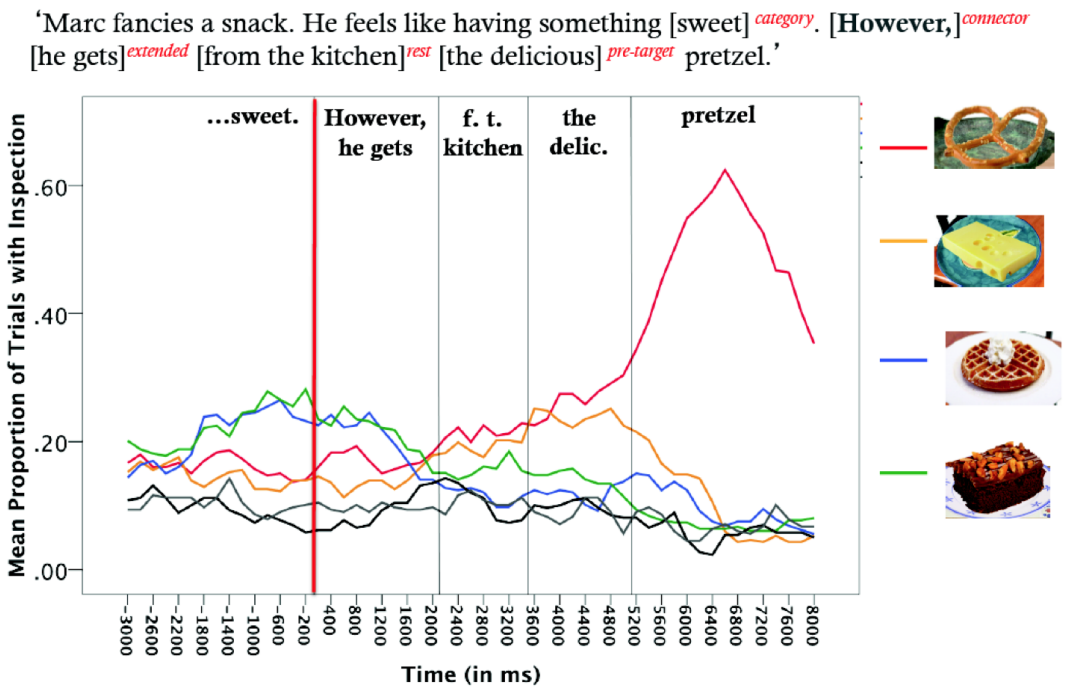


Figure 4.1: Switching gaze after encountering a concessive connective (Köhne and Demberg, 2013)

(5) a. Heidi could imagine and create things because she won first prize at the art show. (Subjective reason)

b. Heidi felt very proud and happy because she won first prize at the art show. (Objective reason)

They compare reading patterns between the two types of relations, which are known to have different degrees of difficulty, and find that increased reading time for the more difficult relation (diagnostic) shows up well in advance to the wrap-up region. Traxler et al.'s finding is supportive of the second view towards comprehension of discourse relations, which posits that the integration of the second clause to the first clause should happen incrementally. According to this account, the first argument of the relation generates some expectation for how discourse will be continued, and this expectation is actively involved in word-by-word processing of the second argument. More recent evidence for incremental discourse processing comes from eye-tracked reading and visual world experiments (Köhne and Demberg, 2013; Rohde and Horton, 2014), as well as, EEG studies (Nieuwland and Van Berkum, 2006; Kuperberg et al., 2011; Drenhaus et al., 2014).

Regarding the effect of discourse connectives on prediction, Köhne and Demberg (2013) ran two eye-tracking experiments on German, one within a visual world paradigm and the other with normal reading setting. The visual word experiment reveals that in a sufficiently constraining context, both causal and concessive connectives generate expectations that direct visual attention to specific objects in a scene before its name is encountered in the audio stimuli (see Figure 4.1). In particular, people look more at an object (e.g., cake) that is congruent with characteristics explained in the preceding context (e.g., "Mary wants to have something sweet") when a causal connective (e.g., *therefore*) is used between clauses. On the other hand, subjects switch to looking more at other objects in the scene immediately after reading a concessive connective (e.g., *however*) which signals an unexpected outcome. These effects show up quite early in the course of listening to the story (right after the connective is encountered) and are indicative of predicting the category of an object before it is mentioned in the audio. The reading experiment revealed prediction effects but only for causal connectives: reading time of the gender-marked region preceding the expected word is shorter than reading time of the same region for the unexpected word given the causal connective. Drenhaus et al. (2014) ran two EEG experiments with similar stimuli in German and English. Both causal and concessive connectives in the German stimuli affect processing of the critical region. A bigger N400-like effect is observed on the pre-nominal region (adjective with gender marking) for the incongruent condition compared to the congruent condition. Hence, the German experiment indicates that readers were able to predict the gender of the upcoming word and hypothetically the word itself, whereas, the English experiment only provides additional support for incremental integration and not prediction of the words. The N400 effect is found on the noun itself not earlier. In addition, both experiments reveal a P600 effect at the discourse connective region in concessive conditions. Based on previous EEG studies, Drenhaus et al. interpret the late positivity as an indicator of the cost of prediction errors and conclude that the concessive connective triggers an update of expectation for a causal relation. These findings suggest that markers of forward causal and concessive relations as two categories among many are processed differently. Not only local integration of the new content is influenced by the connective used to relate a sentence to its preceding context, but also prediction of the critical words becomes possible if stories are designed in a highly constraining manner.

4.2.3 Inference

Linguistic inferences are in the center of pragmatic studies. Inference is a language comprehension process that is involved with propositional units, i.e., full clauses or sentences, unlike the integration and prediction processes which function at various levels like at the word or phrasal levels. Different categories of inference have been discussed in the literature, but *entailment* and *implicature (implication)* are most relevant to our study of discourse connectives. Entailments and implicatures are the product of inference processes on the linguistic input plus world knowledge and they contain statements that are distinguished from or are additional to what is literally said.

Entailment applies when uttering a statement A makes the comprehender directly infer another statement B. In terms of truth-conditional semantics, if A is true then B is also true. For example, (6-a) entails (6-b).

(6) a. There is a grocery shop around the corner.
b. There is a shop around the corner.
c. It is open.

Implication, on the other hand, occurs when a statement C is only suggested by an utterance A, rather than being directly said. For example, if someone answers with (6-a) when they are asked for addressing a place to buy grocery, it would be very likely that the listener infers (6-c) pragmatically.

While an entailment is the product of a properly truth-conditional semantic inference, an implicature is shaped with regard to the cooperative communication principles, e.g., by assuming that the speaker is saying something relevant to the situation (Grice, 1975). An identical utterance can have different implicatures in different contexts and the truth of the implicature is independent of the the truth of the utterance. Without any contradiction, an implicature can be denied by the speaker as the discourse continues (7-a), whereas in case of entailment such a denial would sound odd (7-b). [2]

(7) a. There is a grocery shop around the corner but it's not open.
b. ? There is a grocery shop around the corner but there is no shop around the

[2]Interested readers are referred to a discussion on *"The top 10 misconceptions about implicature"* (Bach, 2006), which includes more clarification on the distinction between entailment and implication.

corner.

One function of discourse connectives is triggering or canceling specific inferences on top of the explicitly uttered arguments. In (7-a), *but* is used to cancel an implication made by the first argument of the relation. This is what makes the entire but-clause relevant to the discourse (Blakemore, 1992). An alternative analysis explains the relation between the two sentences in (7-a) as follows: the first argument implicates something (e.g., "you get grocery.") that an implication of the second argument contradicts. Either view explains the function of the connective with respect to the inferences involved with the two sentences.

A lot of effort in pragmatic theories has been spent on analyzing discourse connectives within an inference-based framework (most recent discussions: Schiffrin (2001); Blakemore (2002); Lewis (2006); Hall (2007)). One controversial topic is whether connectives carry truth-conditional meaning, i.e., whether they add anything to the set of statements entailed by the sentences they combine. Some researchers distinguish between connectives such as *because* and , *before* that affect the truth-conditional state of what is said, and connectives regarded as non-truth conditional, such as *but* and *furthermore*. According to this categorization, an utterance like (8-a) entails (8-d) in addition to the entailments from the individual clauses ((8-b) and (8-c)), whereas a non-truth conditional connective like *but* does not extend the set of entailments with any additional statements.

(8) a. Mary voted for Jerry, because/but John skipped the meeting.
b. Mary voted for Jerry.
c. John skipped the meeting.
d. The reason why Mary voted for Jerry was that John skipped the meeting. (entailed by the use of *because*)
e. John's action was different from Mary's. (implicated by the use of *but*)

Grice introduces the notion of *conventional implicature*, which applies to the case of *but* in (8-a): Given the conventional meaning of *but*, a proposition like (8-d) or other statements can be implicated (in addition to the direct entailments ((8-b) and (8-c)). As Bach (1999) clarifies, conventional implicature, by definititon, is a proposition which is conveyed due to the presence of a certain term with a certain meaning but whose falsity is compatible with the truth of the utterance. Therefore, the difference between the first sentence utilizing

because vs. *but* is that the truth of it when *because* is used depends on the truth of (8-d), whereas in case of *but* the truth of the sentence does not depend on the truth of (8-e).[3]

Unfortunately, we do not know much about the cognitive reality of the above theories. In other words, the experimental data available on processing discourse connectives does not compare with the amount of theoretical discussions on inference. What the so called *conventional meaning* of a discourse connective is, and how it affects the discussed inferences are two important questions that need to be answered experimentally. The aim of this chapter is to show that the distributional account proposed in Chapter 3 can provide a framework to answer these questions. Before we move on, we take another look at the psycholinguistic experiments looking into the inference processes underlying integration and prediction of discourse segments during online comprehension, as well as, the resulting interpretations.

Millis et al. (1995) show that the connective *because* triggers causal inferences that in the absence of the connective would not be inferred by the readers. Subjects were asked to answer verification questions regarding a hypothetically inferred statement after pairs of sentences without a connective, or with *because*, *and*, or *after* were presented in a word-by-word manner. Readers incorporated causal inferences in the presence of *because*, and to lesser extent when *and* was used. The *after* condition did not elicit causal inferences, which further indicates that the effect of *because* was not due to the temporal relation it encoded, rather it related to the specific meaning of *because*. Moreover, the no-connective conditions failed to generate causal inference which in turn is a support for the role of the connectives on inferential processes. In line with their previous finding regarding the *delayed integration* hypothesis (Millis and Just, 1994), reading time of the *because* sentences were longer than their no-connective counterparts, which according to the authors indicates that the inference process consumes cognitive resources but would better sustain content in memory. Some evidence for enhanced memory for sentences connected by causal connectives comes from Caron et al. (1988)'s study, which has been interpreted as a result of inferences that the connective brings about. On the other hand, Murray (1995) found better recall of content for adversative connectives and not for the casuals in their experiment. This raises an awareness regarding the function of individual connective types rather than such coarse-grained categorizations which lead to confusion in interpretation of the findings from different studies. It sounds like the inferences caused by some connectives (from each category) lead to slower processing of the text but instead a better representation of the events in memory, which in turn increases the recall accuracy.

[3]While the coherence or acceptability of (8-d) is affected by the truth of (8-e).

Regarding the time-course of inferences, in terms of putting the two arguments of a relation together and obtaining a new statement, Noordman and Vonk (1992) run a reading experiment with stimuli from unfamiliar topics. Subjects are asked to verify a statement (9-b) that is hypothetically implied from a *because* sentence (9-a) in the text. The experiment consists of two conditions, one of which includes the verification statement in the text prior to the *because* sentence and one doesn't. Noordman and Vonk (1992) hypothesize that if the causal connective is used for a knowledge-based inference in the course of reading, then reaction time to the verification question should be equal across conditions. On the contrary, Noordman and Vonk (1992) find a longer verification time for the condition excluding the explicit statement. This means that the subjects performed the knowledge-based inference marked by *because* only when they were asked to answer the verification question, and not in the course of reading.

(9) a. Chlorine compounds make good propellants because they react with almost no other substances.
b. Propellants must not combine with the product in the spray can.

Note that in this experiment subjects were asked to read the sentences in their normal way without performing an extra task. In a second experiment, Noordman and Vonk (1992) asked people to find inconsistencies in the text they had to read. This time reading time of the second clause in the *because* sentence is significantly different between conditions and no difference is observed in the verification time. This result indicates that people inferred the causal relation during reading and were able to verify the resulting statement quickly and regardless of an overt mention in the text.

These findings support a view that distinguishes between the integration and the inference processes triggered by discourse connectives, despite of the fact that in the majority of studies the two phenomena are considered as a single process. In particular, the incrementality of relational inferences depend on the type of stimuli and the task setup. Remeber Kintsch and Van Dijk's definition of language comprehension at the beginning of this section. If comprehension of a discourse relation is a process leading to the subject's ability to answer questions about not only the directly uttered sentences but also the inferred statements, then comprehension does not occur only by integration. Cozijn et al. (2011) define integration as the process of finding the relation between the connected sentences with the help of the text-internal devices such as discourse connectives and referential devices. On the other hand, world-knowledge inference, according to Cozijn

et al. (2011), refers to the process of deriving the general causal relation and checking it against the comprehender's world knowledge. In an eye-tracking experiment, Cozijn et al. (2011) exposed the reader with short narrative text including a causal relation either with or without the connective *omdat*, which is the Dutch equivalent of *because*. Comprehension questions were designed to make sure that subjects read the text carefully: an implied statement like (10-b) is displayed and people should decide whether it is true or false.

(10) a. ... On his way to work he experienced a long delay, because there was a large traffic jam on the highway.
b. A traffic jam leads to a delay.

Reading times on the region immediately following the connective, middle region of the subordinate clause and the wrap-up region were analyzed separately. Reading times of the first two regions were smaller in the connective-present condition, whereas the wrap-up region was processed more slowly compared to the no-connective condition. This result is similar to previous findings of Millis and Just (1994), while a finer analysis is undertaken. A significant interaction of region and conjunction is observed and interpreted as an evidence for the differential time courses of the integration vs. inference processes. This means that Cozijn et al. (2011) try to distinguish between the type of processes occurring incrementally and the type of processes involved when the subject reaches the end of the sentence. The former is indicative of shallow integration and the latter is a deep inferential process. A second experiment on the same material, this time in a self-paced reading framework with a time-recorded verification task, was conducted to look into the inferences made offline: people were asked to verify the statement as quickly as possible and try to answer correctly. Again in the connective-present condition, the total reading time of the second clause did not differ, but the processing of the middle part of the sentence was shorter and the final region was read slower. Moreover, readers were faster in the verification of the inferred statement when the connective was present. Therefore, the authors conclude that while integration in terms of connecting the second argument of the relation to the first is an immediate effect of the connective leading to faster reading of the middle part of the second clause, inference takes place when the propositional content of the second clause is accessible, i.e., at the end of the story. Furthermore, the presence of the connective makes it easier for people to compare the result of that inference against their world knowledge.

There is one point regarding the stimuli in both experiments performed by Cozijn

et al. (2011) which should be noted when talking about making inferences triggered by discourse connectives. The verification statement in this experiment is part of the common knowledge, it is not an additional statement that is derived about the state of affairs in the story. Recall our example of the entailment obtained from the use of *because* in a sentence like "Mary insisted that they should meet, because John disagreed.": The reason why Mary insisted was that John disagreed. This statement that is made explicit by the connective together with many other implications that are not explicit in the text are inferred even if we have no real-world knowledge about Mary and John. Since the context (the two relational arguments) are neutral with regard to the type of relations they can have, the connective takes over to guide the listener to select the right interpretation (Sanders and Noordman, 2000). At the same time, world-knowlege is involved to determine the acceptability of the relation cued by the connective. Connectives in the wrong places, i.e, incompatible with the set of possible relations the two arguments can have, cause processing difficulty. Experiments by Köhne and Demberg (2013) and Drenhaus et al. (2014), which we reviewed in the previous subsection, showed that during online reading people were able to predict how the second argument of *therefore* vs. *however* (and the German equivalents) should be continued content-wise, and reacted immediately to the unexpected second clause in the incoherent conditions. This can be taken as evidence for the world-knowledge inferences taking place quite early, before the complete content of the second argument is available. Nevertheless, given that various types of inferences are possible, more research is needed to elaborate the differential and cooperative effects of the clausal content, world-knowledge and that of linguistic marking with the help of explicit connectives. In fact, some inferences are the result of the content of the sentences put together, as they construct a specific relation, and some are drawn from the meaning of the connective. Sanders and Noordman (2000) elaborate on the each of these factors and argue that the relations between the sentences are the basis for understanding language comprehension at the level of discourse. According to Sanders and Noordman, relational markers have effect during online processing but their influence decreases over time, whereas the effect of the coherence relation is robust and determines the mental representation of the text. The idea that distinguishes Sanders and Noordman (2000)'s approach from previous theories is that discourse relations are the basis for understanding the effect of discourse markers, not the other way around. Relations are the subject matter in a cognitive study of discourse processing, whereas connectives are linguistic devices to guide those processes. This idea appeals to our distributional representation of the connective meaning with regard to the discourse relations it can help to infer.

4.2.4 Open questions

The review in this section indicates that connectives have been repeatedly found to facilitate linking sentences to one another at the surface level. When it comes to deeper comprehension, i.e., making inferences based on the world-knowledge relevant to the content of the involved propositions, it is still unclear to what extent the connective's specific meaning would influence the interpretation.

Causal connectives have obtained particular attention in experimental studies on linguistic inference (among many others, (Noordman and Vonk, 1992; Millis et al., 1995; Traxler et al., 1997; Sanders and Noordman, 2000; Cozijn et al., 2011)). Findings of these studies indicate that people benefit from the presence of a connective like *because* to activate their world-knowledge about specific events mentioned in the text and compare it against the content of the sentences. The concessive connectives known as the negative counterparts of causal connectives (König, 1991) have been less explored. A few recent studies suggest that, during online comprehension, causal and concessive connectives are not processed exactly in the same way (Köhne and Demberg, 2013; Drenhaus et al., 2014; Xiang and Kuperberg, 2014; Xu et al., 2015). These experiments have been conducted with two typical designs: two sentences connected with/without a certain connective, or sentences linked by discourse connectives from very different categories, e.g., causal *therefore* vs. concessive *nevertheless*. The causal or no-connective condition is taken as a baseline for understanding how concessive relations affect inferences. This setup gives us some general view of the way counter-factual relations are anticipated when concessive discourse markers are used but does not elaborate on the type of inferences involved. One question that needs to be answered is whether certain extra statements are inferred from a pair of sentences when they are linked by particular discourse connectives. Such inferences might influence processing of the larger context like an upcoming sentence in the text, and this is left unexplored in previous work.

A second question is involved with the fine-grained inferences triggered by connectives of similar types. For example, when the linguistic context allows for using *but* and *although*, it would be interesting to see what different implications might pop up in the presence of either connective. Questions of this type have involved many theoretical linguists for decades, but no experimental study has been conducted to resolve controversies on the different procedures led by similar connective types.

Finally, the amount of information delivered by a connective to distinguish one relation from another has been overlooked in previous work. As we saw in our general

analysis of PDTB, while some connectives are strong markers of very specific discourse relations, some others like *but* and *and* are used in a variety of discourse relations. The question is how far the information content of the connective affects comprehension processes. This question is concerned with a quantitative rather than a qualitative aspect of the experimental studies.

To remind us about the main goal of this chapter, we want to examine some predictions we make in the framework of an information theoretic account of discourse connectives against previously developed ideas on the way individual connective types affect text comprehension. Our theory suggests that connectives occurring in multiple relation senses deliver some information about each such relation. The more frequently a connective co-occurs with a relation, the more bias towards that interpretation when the connective is used in a new context. In order to make clear how this perspective could explain experimental findings, remember traxler1997processing's study of simple causal vs. diagnostic relations. The processing difficulty in comprehension of the diagnostic relations compared with the reason baseline was attributed to reader's effort to construct a mental space in which the real consequence of an event becomes its evidence. The distributional account, however, explains the difficulty attached with processing of the diagnostic relations by considering the strength of the discourse marker *because* for the diagnostic vs. reason relations. Since in natural distribution of discourse relations in English, *because* is most frequently used in reason relations, sentences compatible with a reason interpretation are processed faster when connected by *because*. In case of diagnostic relations, *because* is not a perfect marker, therefore subjects should look for the coherence relation on their own, i.e., with minimal help from the connective. The facilitating effect of the other linguistic markers of diagnostic relations can also be explained following the same intuition. It is quite possible that in some other language, diagnostic relations have a more specific discourse connective (something that should perhaps be translated to *given that* in English), which distinguishes these relations from simple causal relations and consequently facilitates processing of the diagnostic relations the same way *because* facilitates processing of the reason relations. The experiment by Canestrelli et al. (2013) on the Dutch causal connectives *want* and *omdat* indeed proves this point. They find that knowledge of the usage patterns attached to each connective helps language users during reading. The subjective connective *want* (typical marker of diagnostic relations) induces a subjective representation, whereas the objective connective *omdat* (typical marker of semantic reason relations) triggers an objective representation. Using either of the connectives in the relation they don't typically mark leads to additional processing difficulty.

Considering the open questions mentioned above, I have selected two English concessive connectives, *but* and *although*, from PDTB for an empirical examination. As we see in the next section, the two connectives are used in similar sets of discourse relations and, according to previous theories, should be involved with closely related inference processes. However, detailed differences in their distribution across discourse relations of various types would shape differential hypotheses regarding their effect on interpretation of identical stories. I show that these hypotheses are more accurate than those which ignore the effect of frequency and information content of the connectives.

4.3 Approaching multi-sense connectives

In this section, I focus on the differences and similarities between *but* and *although* as two frequently used connectives in adversative relations. An additional feature that makes this couple an interesting case for a comparative study is their ambiguity. The predictive power of the information theoretic account employing a distributional meaning representation can be highlighted when dealing with multi-sense connectives.

4.3.1 Distribution of *but* in PDTB

There is a total of 3308 cases of *but* annotated in Penn Discourse Treebank with either a single discourse relation sense label (3276 instances) or two (32 instances). In all analyses in this thesis, the double-tagged connectives are counted for each relation sense separately, unless otherwise is explicitly indicated. Therefore, our target set of relations using *but* as their connectives includes a total of 3340 relation instances. Senses that occurred more than 10 times in this set of relations are presented in table 4.1.

At this point we need to learn about the definition of each of these relations in the PDTB terminology. Definitions are copied from the original annotation manual (Prasad et al., 2008). We start by the high-level class of relations, i.e. `COMPARISON`, which is equivalent to the set of relations that I so far called adversative, or negative-polarity relations:

> The class tag `COMPARISON` applies when the connective indicates that a discourse relation is established between Arg1 and Arg2 in order to highlight prominent differences between the two situations. Semantically, the truth of both arguments is independent of the connective or the established relation.

The class `COMPARISON` has some more specific subtypes. However, as the table shows,

Table 4.1: Relation senses annotated for the occurrences of *but* in PDTB

Original label	Frequency	Percentage in *but* cases
COMPARISON.Contrast	1612	48.3
COMPARISON.Contrast.juxtaposition	639	19.1
COMPARISON.Concession.contra-expectation	496	14.9
COMPARISON	263	7.9
COMPARISON.Contrast.opposition	176	5.3
EXPANSION.Conjunction	86	2.6
COMPARISON.Pragmatic contrast	30	0.9
COMPARISON.Concession.expectation	12	0.4

about 8 percent of the *but* relations have been left unspecified. It means that the annotators did not find these relations compatible with any of the more specific definitions. Examples of the `COMPARISON` annotations are (11). Some of these cases have other discourse cues which might have confused the annotators to decide which specific relation sense applied. Others seem to be speech-act relations that are not considered in the PDTB relation hierarchy (d,e).

(11) a. I'm for the Giants today, but **only because** they lost yesterday. (22:2)
b. Retail profit surged, but the company said it was **only** an "odest contributor" to third-quarter results. (22:19)
c. He visited the Hugo devastation but **not until** after local leaders urged him to do so. (19:20)
d. The market can adjust to good news or bad news, but uncertainty drives people wild. (14:18)
e. You can find some good, quality companies over the counter, but investors should be selective. (14:40)

Now let's have a look at the most frequent sense of *but*, that is the `COMPARISON.Contrast` relation:

> `Contrast` applies when the connective indicates that Arg1 and Arg2 share a predicate or property and a difference is highlighted with respect to the values assigned to the shared property. In `Contrast`, neither argument describes a situation that is asserted on the basis of the other one. In this sense, there is no directionality in the interpretation of the arguments.

`Contrast` relation has been annotated for about half of the occurrences of *but* in the corpus, including the following examples.

(12) a. They're trying to plug the various loopholes, but they're totally unprepared for this. (6:29)
b. Futures prices rose modestly, but trading volume wasn't very heavy. (10:26)
c. The carnage among takeover stocks Friday doesn't mean the end of mega-mergers but simply marks the start of a less ambitious game. (24:43)

The more specific sub-types of `Contrast` are `juxtaposition` and `opposition`, which are closely related and in most previous studies have been referred to as opposition:

> The subtype `juxtaposition` applies when the connective indicates that the values assigned to some shared property are taken to be alternatives. More than one shared predicate or property may be juxtaposed. The subtype `opposition` applies when the connective indicates that the values assigned to some shared property are the extremes of a gradable scale, e.g., tall-short, accept-reject etc.

Examples of `juxtaposition` and `opposition` are shown in (13-a) and (13-b) respectively.

(13) a. That may be the largest patent award ever, but it is well below the $12 billion Polaroid seeks. (21:54)
b. Terminals at San Francisco International also were damaged, but the tower itself was intact. (18:3)

Finally, about 15% of the *but* occurrences are annotated with `Concession` relations which are explicitly defined based on an underlying causal inference.

> The type `Concession` applies when the connective indicates that one of the arguments describes a situation A which causes C, while the other asserts (or implies) C'. Alternatively, one argument denotes a fact that triggers a set of potential consequences, while the other denies one or more of them.

The PDTB manual includes a formal representation of the `Concession` relation, which might have been adopted from previous work (König, 1991). According to König, "P but Q" in this sense implicates $P \rightarrow R$ and $Q \rightarrow R'$, where R' can either be a hidden statement

not mentioned explicitly in the text, or it could be equivalent to Q.[4] The difference between the two subtypes of `Concession` is a very important one, as it relates to the direction of causal inference or dependence between the two arguments.

> Two `Concession` subtypes are defined in terms of the argument creating an expectation and the one denying it. Specifically, when Arg2 creates an expectation that Arg1 denies ($A = ||Arg2||$ and $B = ||Arg1||$), it is tagged as `expectation`. When Arg1 creates an expectation that Arg2 denies ($A = ||Arg1||$ and $B = ||Arg2||$), it is tagged as `contra-expectation`.

As one can see from the table, *but* appears frequently in only one of the two sub-types of the `Concession`, that is the `contra-expectation` (14). From now on, when talking about concessive relations, I always refer to the underlying causal inference by "$P \rightarrow Q$" and to the argument violating this causality as demonstrator of Q' (which is Arg1 in `expectation` and Arg2 in `contra-expectation`).

(14) a. Korean car exports have slid about 40% so far this year, but auto makers here aren't panicking. (21:40)
b. Avondale asked Travelers to defend it in the state proceeding, but the insurer didn't respond. (18:19)
c. Mr. Baker will relinquish his previous positions, but a successor for him hasn't been named yet. (12:38)

Very few instances of *but* are annotated with the `expectation`. In some of them, the causality is noticeable (15-a), but some others sound to be very similar to `Contrast` relations (15-b).

(15) a. You might find something, but the chances are low. (21:54)
b. A&W Brands lost 1/4 to 27. But its third-quarter earnings rose to 26 cents a share from 18 cents a share last year. (20:19)

There are instances of *but* annotated with `Pragmatic contrast` relation which is defined as follows:

[4] In the work of Konig, however, only the latter (the more specific condition with $P \rightarrow Q$ where one argument states Q') is called concession.

The tag `Pragmatic contrast` applies when the connective indicates a contrast between one of the arguments and an inference that can be drawn from the other, in many cases at the speech act level.

As it is described in the definition and can be seen in the examples, there is no contrast between the situations described in the two relational arguments. In fact, in half of the *but* cases annotated with `Pragmatic contrast`, a second label has also been given, which in most cases is `Expansion.Conjunction` (16).

(16) a. We ran into the house to get Mame, but the next tremor threw me in the air and bounced me as I tried to get to my feet. (17:78)
b. DPC made a $15-a-share bid for the company in May, but Dataproducts management considered the $283.7 million proposal unacceptable. (6:97)
c. Mr. Engelken moved south to Washington, but he took with him enduring memories of the homer of 1951. (7:58)

Now that we saw examples of *but* usages in a veraiety of discourse relations, it is time to have a look at a more specific connective *although*.

4.3.2 Distribution of *although* in PDTB

A look at the table of *although* statistics in PDTB (Table 4.2) indicates that this connective applies to most of the relation senses that we found for *but*. There is an obvious difference though, between the distribution of *although* and *but* across these relations.

Table 4.2: Relation senses annotated for the occurrences of *although* in PDTB

Original label	Freq	Percentage
COMPARISON.Concession.expectation	132	40.2
COMPARISON.Contrast	114	34.8
COMPARISON.Contrast.juxtaposition	34	10.4
COMPARISON.Concession.contra-expectation	21	6.4
COMPARISON	16	4.9
COMPARISON.Contrast.opposition	9	2.7

A total of 328 ocurrences of *although* has been annotated in PDTB, among which no double-tagged relation is labeled. `Expectation` which was at the bottom of the table

for *but*, takes over the first place as the most frequent relation annotated for *although*.

(17) a. Although that may sound like an arcane maneuver of little interest outside Washington, it would set off a political earthquake. (6:9)
b. Although working for U.S. intelligence, Mr. Noriega was hardly helping the U.S. exclusively. (20:13)
c. Third, oil prices haven't declined although supply has been increasing. (2:31)

On the other hand, the three most frequent senses of *but*, i.e., `Contrast`, `juxtaposition` and `contra-expectation` fall behind. Also, the underspecified label of `COMPARISON` has been annotated only for about 5 percent of the *although* relations. At this point, I would like to make an observation regarding the statistics of the mid-sentence usage of *although* (124 instances) vs. the sentence initial occurrences of this connective (204 instances). Table 4.3 separates between the two arrangements of *although*. We notice that the distribution of the mid-sentence *although* over relations of different types drastically deviates from what we saw in the total occurrences. There is a significant correlation between the arrangement of *although* and the annotated relation sense ($\chi 2 = 76.2183, df = 7, p - value = 8.111e - 14$).

Table 4.3: Distribution of the relation senses annotated for occurrences of sentence-initial vs. mid-sentence *although*

Original label	Initial	Perc.	Middle	Perc.
COMPARISON.Concession.expectation	**115**	**56.4**	17	13.7
COMPARISON.Contrast	60	29.4	54	43.5
COMPARISON.Contrast.juxtaposition	17	8.3	17	13.7
COMPARISON	5	2.5	11	8.9
COMPARISON.Contrast.opposition	4	2.0	5	4.0
COMPARISON.Concession.contra-expectation	2	1.0	**19**	**15.3**

The distribution of relations marked by mid-sentence *although* is more similar to the distribution of *but*. Not only we see more tendency towards marking contrastive relations, but even among the concessive types, the mid-sentence *although* shows a significant tendency for `contra-expectation`, a relation sense that appears only one percent of the time with sentence-initial *although*. Remember the difference between the the two subtypes of `Concession`: In `contra-expectation`, which appears often with *but*, and apparently with mid-sentence *although*, the clause introduced by the connective is

an unexpected statement given the implication generation by the other clause. Examples (18-a) and (18-b) are of relations which dominantly co-occur with *but* and then with medial *although*, that are `Contrast` and `contra-expectation`, respectively. They indicate a specific similarity between *but* and this arrangement of *although*.

(18) a. It won't increase its offer although adjustments within the proposed pay-and-benefit mix are possible. (19:18)
b. Ms. Levine had never been fired, although she had stopped working at the restaurant. (16:92)

As a summary, the distribution of *but* and *although* indicates some similarities and some differences between the two connectives. In the next section we use this information for hypothesizing about the effect of each connective on inferential processes that help English speakers interpret connected sentences in a certain way, where different interpretations are possible.

4.3.3 Comparing the two connectives

According to the framework I proposed in Chapter 2, the cognitive effect of a discourse marker has to be determined with respect to the information it delivers. This information is equal to the amount of uncertainty the connective removes by shrinking the search space of the discourse relations that might be inferred from the combination of the two sentences. A major difference between *but* and *although* is visible in the distribution of the second-level relation senses, that is the tendency of *but* to mark `Contrast` relations vs. that of *although* to mark `Concession` types. What can we infer from this finding about the way *but* and *although* affect sentence comprehension? Assuming that the context, i.e., the content of the two relational arguments, gives the reader freedom to interpret the relation as either of the `COMPARISON` types, would *but* bias interpretation towards `Contrast` and *although* towards `Concession`? Would the finer grained differences (frequency of `expectation` vs. `contra-expectation` in the distributional representation of each connective) matter at all? To answer these questions will need to design experiments in which readers are implicitly examined for the interpretations they make. In this section, I first discuss a property of discourse relations which takes different values across `COMPARISON` types. This property deals with the notion of hierarchical semantic connections between sentences (Blühdorn, 2008) and is implicit in the definition of relation

senses. We then determine the value each type of COMPARISON takes at this dimension and use this comparison to predict how differently the two connectives *but* and *although* should affect interpretation of the sentences.

The semantic asymmetry dimension: In the theory of syntax, one way to establish hierarchical connections between clauses is using a subordinating conjunction, such as *because* and *although*. A subordinate clause is not a complete sentence and in some languages like German, its structure is affected by the connective (e.g., the verb is moved to the end of the sentence). Therefore, the structure of the introduced clause by a subordinate conjunction is, namely, *governed* by the main clause. According to Blühdorn (2008), semantic hierarchy is instead defined based on the (a)symmetry of the relation between two semantic units. If the conjoint units have **equal semantic functions** and **equal semantic weights** the relation between them is symmetric or non-hierarchical, otherwise it is asymmetric or hierarchical. When applied to semantic relations between clausal units of text we find a variety of both types of relations. Blühdorn (2008) exemplifies the following ones:

(19) a. The penguins were yellow-brown, and the giraffes were black and white. (Symmetric)
b. Mary went to the library, and she began to feel hungry. (Asymmetric)

One of the syntactic consequences of semantic symmetry, according to Blühdorn (2008), is the possibility of exchanging the relational arguments without a significant change of meaning. On the other hand, in hierarchical semantic connections the meaning changes by reordering the arguments.

(20) a. The penguins were yellow-brown, and the giraffes were black and white.
b. The giraffes were black and white, and the penguins were yellow-brown.

(21) a. Mary went to the library, and she began to feel hungry.
b. Mary began to feel hungry, and she went to the library.

Blühdorn provides evidence for the argument that syntactic and semantic hierarchies are

not equivalents. First of all, each type of hierarchy is shaped by a different set of devices. For example, discourse connectives including sentences conjunctions and adverbials are devices for obtaining semantic hierarchies, whereas syntactic hierarchy can be made by subordinating conjunctions, complementizers, relative pronouns, and infinitives. Secondly, both coordinating and subordinating conjunctions can possibly be used in symmetric and asymmetric semantic relations.

(22) Symmetric relations:

a. The giraffes were black and white, **and** the penguins were yellow-brown.
b. The giraffes were black and white, **while** the penguins were yellow-brown.

(23) Asymmetric relations:

a. Mary went to the library, **and** she began to feel hungry.
b. Mary went to the library, **before** she began to feel hungry.

Semantic asymmetry in *but* vs. *although*: The above definition of semantic hierarchy suggests that symmetry/asymmetry is not a fixed attribute of the connective types, rather should be treated as a characteristic of the discourse relations. In our study, discourse relations are equivalent to semantic relations between clauses. Other schemes like RST (Mann and Thompson, 1987) and SDRT (**?**) propose their own notion of discourse hierarchy, which are also discussed in Blühdorn (2008). Since PDTB relations are considered to be semantic relations, I find Blühdorn's account of semantic hierarchy most suitable and clear for our analysis. Nevertheless, the phenomenon is in many ways related to other analyses of discourse structure and information structure — for example, see Stede (2007)'s discussion on nuclearity in RST relations. It is time to look back into the adversative relations that occur with *but* and *although* in the corpus, and categorize them into symmetric and asymmetric relation senses. In the above discussion, we saw examples of `opposition` (22). By looking into more examples of this relation as well as its sibling, namely `juxtaposition` in PDTB, we find that these two should be categorized as non-hierarchical relations. More precisely, Arg1 and Arg2 in these relations have the same semantic function and same weight. The underspecified `Contrast` relations are difficult to be categorized as hierarchical or not. On the one hand, the general definition of `Contrast` indicates no directionality between the two arguments, on the other hand, the variety in examples annotated with this label in PDTB makes it difficult to say that this

relation is hundred percent symmetric. For now, we just assume that Contrast relations are underspecified with respect to this dimension. Both subtypes of Concession should be considered as hierarchical. At least the semantic function of the two arguments are different according to the definition of these relation senses: Remember that one argument is a premise or expectation generator, while the other is the denier or the unexpected outcome regarding the underlying the "$P \rightarrow Q$" inference. But, how are we to determine the argument with a bigger/smaller semantic weight, or as I call it from now on, the emphasized or the more salient argument? In Concession, the argument demonstrating P is the premise of a violated inference, whereas the other argument demonstrates an unexpected assertion, though, a real state of affairs. I propose that the unexpected assertion is a bigger information update since it does not comply with common knowledge, thus, should be considered as a more salient statement in the discourse compared with the other argument. In expectation relations, this more important statement, representative of Q', corresponds to Arg1, and in contra-expectation relations, it corresponds to Arg2. Therefore, the two relations differ in terms of the argument each of them highlights.

Table 4.4 summarizes argument salience information attached to the three connectives of our interest based on the relations they mark in PDTB. As we found, *but* is used most often in relations which enforce no causal dependence between the two arguments, i.e, Contrast and its subtypes. While we count opposition and juxtaposition as relations with equal salience for both arguments, the under-specified Contrast annotations are left unknown. Among the Concession relations, *but* is observed most often with contra-expectation, in which the argument attached to the connective (Arg2) is salient. According to the distributional hypothesis, the function of *but* should be defined based on the frequency of its occurrences across the relations we discussed. This means that in any context, *but* should have some bias towards highlighting Arg2. On the other hand, *although* in the sentence-initial arrangement occurs most often with the expectation relation, the other subtype of Concession which implies a direction reversed to that of contra-expectation: the clause attached to the connective is a denied expectation. Therefore, the distributional hypothesis predicts a higher likelihood for the other argument (Arg1) to be salient when *although* is used at the beginning of the sentence. Finally, we found that mid-sentence usage of *although* is closer to *but* in terms of distributional similarity. A closer look into the subtypes of the Concession relations, however, suggests that *although* in this arrangement does not show any significant preference as to which of the arguments should be more salient. Thus, a fully distributional account of *but* and *although* predicts that the effect of mid-sentence *although* on interpretation of causal dependencies, direction of the inference and determining the

Table 4.4: Distribution of argument salience in PDTB relations

Salient argument	*but*	Initial *although*	Middle *although*
Arg1	12	**115**	17
Arg2	**496**	2	19
Both	815	21	22
Unknown	2656	83	83

more important argument should be an average of the effect of *but* and the sentence-initial *although*.

In order to test these hypotheses, we need to find pairs of sentences which can be interpreted in various ways, and see how alternating between the connectives and sentence arrangement would favor one interpretation over the others. We saw that all these connectives are similar with respect to the coarse-grained relations distribution. In other words, both arrangements of *although*, as well as *but* occur always in `COMPARISON` relations (with some exceptions for *but* appearing in *Expansion* relations which we discussed previously). Therefore, difference between the inferences triggered by these connectives should be subtle and difficult to measure by online methods. That is why we start with a coherence judgment study. Before proceeding with the experimental design, though, I would like to show how predictions of the classical approaches to connectives' meaning compare with that of the distributional account.

4.3.4 Alternative accounts of *but* and *although*

The classical approach to define the meaning of discourse connective is to focus on a single connective type, find a range of examples in which the connective applies and try to form a set of rules about the type of contexts it can or cannot be utilized. In this section, I only look into the most recent and prominent approaches to formalize the meaning of *but* and *although*.

Fraser (1999) exemplifies the following uses of *but* and proposes a *core meaning* approach according to which all instances of *but* are explained with respect to a single function: *but* marks **simple contrast**.

(24) a. She's good looking. But he's ugly as sin. (opposition)

b. He's good looking. But that isn't going to get him a job in this market. (contra-expectation)
c. You say that Mary is coming. But we weren't talking about Mary at all. (pragmatic contrast)

It is not clear though, what simple contrast means. The term *simple* only implies that a generalization over all functions of *but* is concerned. Based on this assumption, one could map this relation to the COMPARISON class in the PDTB hierarchy. Other types of connectives can also mark some sort of contrast. Fraser (1998) states that these connectives have slightly different core meanings, nevertheless, *but*, as the more general marker of contrast, can substitute them. For example, *however*, and *although* are categorized as connectives very similar to *but*. But then a difference is pointed out regarding the way each connective emphasizes one of the two relational arguments. This difference is not treated as an important factor in Fraser (1998), i.e., part of the definition of contrast, rather it is mentioned to distinguish the more specific connectives *however* and *although* from the more general connective *but*. For the following examples, it is argued that *but* treats both arguments equally, *however* emphasizes on the first clause by putting the second clause in a sub-ordinate position, and *although* places priority on the second clause.

(25) a. She fried the onions, but she steamed the cabbage.
b. She fried the onions, however she steamed the cabbage.
c. She fried the onions, although she steamed the cabbage.

Fraser's distinction does not follow the syntactic subordination perspective, given that *however* and *although* are treated differently. Yet, it is not explained whether his rule should be applicable for all usages of these connectives or only in a specific context. If that is the case, then our predictions for *but* and *although* differs from Fraser at least in contexts where concessive inference is possible: we expect bias towards a contra-expectation relation, thus more emphasis on the second argument of *but*, whereas Fraser predicts a rather neutral effect for this connective. Also, we distinguish between the mid-sentence and initial usage of *although*, but such a distinction is absent from Fraser's analysis.

Search for a unitary meaning definition of *but* is continued by other researchers. Blakemore (2002) argues against Fraser's approach to view *but* as a representative of the generalized contrast. She exemplifies cases of contrast where *but* cannot safely substitute

other connectives. For example, in (26-a), the source of contrast between the two sentences is that one implicates the lady is having though time, while the other implicates that she is having good time. Substitution of *and* with *but* as a marker of contrast should not affect the coherence, but in reality it does. The difficulty is not a loss of meaning due to replacing *and* with a more general marker of contrast, it rather relates to the fact that not all contrast relations are equally coherent with *but*.

(26) a. Her husband is at the hospital and she's dating other men.
b. ?Her husband is at the hospital but she's dating other men.

Blakemore proposes a procedural analysis of *but* according to which discourse connectives are essential guides for the listener to the interpretation that the speaker intends. After analyzing a wide range of *but* sentences, she assigns a universal procedure to this connective, that is, to eliminate an assumption shaped in the reader's mind based on preceding context. The term used to refer to this core function of *but* is **denial of expectation**. Within the general framework sketched by Blakemore for analyzing discourse cues, using *but* in a place where the first argument of the relation does not generate any expectation of the type the second argument removes would violate the principle of relevance (26-b).[5] While in Fraser's attempt to formalize the meaning of *but* all uses are reduced to contrast, Blakemore tries to show that denial of expectation is the universal function of *but*, which presumably covers not only the concessive but also the contrastive uses of the connective.

Almost in all previous theories, *although* has been viewed as a connective with more limited and more specific usage compared to *but*. One of the well-known accounts of concessive connectives is developed by König (1991). In this account concession relations are viewed as the dual of causal relations. The formal representation for a sentence of the form "Although P, Q." is equivalent to that of "Because P, Q'." in the duality account. The underlying causal inference is "If P, normally Q." in both relations. As pointed out by Iten, while this provides a basis for studying concessive use of the connective, defining this relation does not lead to a full account of the meaning of *although*. By taking a similar approach to that of Blakemore in case of *but*, Iten tries to define a universal procedural meaning for *although*: "this connective suspends an inference from what follows, which would result in an unresolvable contradiction." For example, in exposure to (27-a) the

[5] A complete introduction of the relevance-based approach to discourse cues needs an extended background on the theories of communication from Grice (1975) to Wilson and Sperber (2002) that is out of the scope of this chapter.

hearer first processes Q that is the content of Arg1, then *although* indicates that there is an inference from P, i.e., the second argument that has to be suspended. The inference that would yield in a contradiction is of the form $P \rightarrow Q'$.

(27) a. Peter went out although it was raining.
b. Peter went out but it was raining.

Iten (2000) explains that the above definition applies also to cases where the causal inference is indirect. Remember the general definition of `Concession` in PDTB, which covers cases where Arg1 and Arg2 state P and R whose implications are contradictory: $P \rightarrow Q$ and $R \rightarrow Q'$. These cases do not fit into König's concession relations,[6] but are compatible with the definition of *although* proposed by Iten.

(28) a. I need some fresh air although it's raining.
b. Although it's raining, I need some fresh air.

The `Contrast` relations formulated by *although* are exemplified by (Iten, 2000) also as reducible to denial of expectation, but she does not thoroughly explain them after she proposes the universal procedure of the connective.

(29) John is tall but Bill is short.

Hall (2004) focuses on this frequent usage of *but* (which according to Iten (2000) should be possible to make with *although* as well) and argues that neither a denial nor an expectation is involved. Following Lakoff (1971), he states that a sentence like (24-a) only draws attention to the fact that the two people contrast with respect to a specific attribute. It is possible that the relation appears in some context that could be interpreted as a denial of expectation, but not necessarily. Therefore, a correction to previous accounts is proposed by Hall: "What *but* is doing in the obvious denial cases is indicating that the hearer is not to draw a conclusion that he could be expected to draw. The more contrast-like cases, among others, show that it can't necessarily be an inference that the hearer was expected to make that is getting cut off by the but-clause; however, it does seem that there has to be an

[6] Both König (1991) and Iten (2000) call them adversative relations.

at least **potential inferential route** that is cut off for the use of but to be acceptable. ". In simpler words, *but* according to this account suspends an inference that would result in a contradiction with what follows, and there is a continuum of cases where such an inference is involved. Moreover, Hall (2004) explains that the function he is proposing for *but* is in principle same as what Iten (2000) proposes for *although*, except that, given the underlying defisible causal rule $P \rightarrow Q'$, *although* introduces P, whereas *but* introduces Q. Therefore, in a context where either of the involved clauses could be denied, *but* and *although* should have differential effects when directly replaced with one another. In other words, "X but Y." denies X and "X although Y." denies Y. According to this account, the difference between "X although Y." and "Although Y, X." is rather undetermined given that in the first place the unitary meaning of *although* should be kept across usages. Iten (2000), nevertheless, dedicates a section to explaining in what context the two arrangements of *although* should be different. She distinguishes between the direct and indirect concessive uses of *although*, and states that in direct concession (30), no difference exists in acceptability of the two possible arrangements.

(30) a. Peter went out although it was raining.
b. Although it was raining, Peter went out.

However, when it comes to indirect concession (31), where the suspended inference is from Arg2 to the negation of an implicature of Arg1, the two versions are different with respect to the processing effort involved with the inference.

(31) a. He has long legs although he is a bit short of breath.
b. Although he is a bit short of breath, he has long legs.

In particular, Iten suggests that the fronted version of *although* is more acceptable because the other version involves a complicated path of inference: the underlying causal inference, i.e., "If X is short of breath, X is not a good runner." is only possible to be made after the entire sentence is read, and then the first clause's content needs to be incorporated. Therefore, the difference between the two usages of the connective pointed out by Iten (2000) does not have anything to do with various interpretations, it rather relates to processing difficulty. Moreover, no particular context is exemplified in which the mid-sentence *although* is considered to be the more coherent formulation.

Studies reviewed in this section starting from the very simplistic account of Fraser (1999) to the more detailed relevance-based approaches (Iten, 2000; Blakemore, 2002; Hall, 2004)all have one thing in common: for every connective a unique and general meaning is being sought. Our corpus study of *but* and *although* suggests that the general relation which covers all meanings of each connective is `COMPARISON`. The name label is not as important as the fact the the same general relation is obtained for both connectives. This proves that a unitary-meaning approach would not be adequate if we want to compare the effect of two connectives with overlapping usages. In such an approach, the difference between two connectives in terms of their frequency, or degree of specificity for marking different types and sub-types of a general relation would get lost. In our framework, a connective's function is rather defined based on its information content and the information content of a connective is obtained from its probabilistic distribution across relation senses. From this perspective, *but* and *although* each should bias interpretation towards the relation that it often co-occurs with. We focused on the semantic hierarchy dimension, and based on the distribution of relations in PDTB, hypothesized for each connective type which of its arguments should become salient in discourse. This is, basically, a testable rewrite of **what relation is more likely to be inferred when either connective is encountered by the reader**. In previous theories we did not find a concrete answer for this question. Except the syntactic subordination account and a few places where a side note is made by previous pragmaticians on usages of *but* and *although*, we do not know much about the emphasis put on the content of a sentence when a particular discourse connective links it to the preceding/following context. Experiments in the following sections examine our hypothetical answer based on the distributional study to the above questions, as well as the more general question of the chapter, that is **whether the distributional representation of a connective is indicative of its comprehension effects**.

4.4 Experiment 1: *but* vs. *although* in identical context

In this experiment we examine how using either *but* or *although* in concessive context affects the interpretation of a relation and whether the results we get from a coherence judgment task would be compatible with predictions of the corpus study. If readers come up with an interpretation of the story which involves a causal inference, then we expect the following effects:

- *But* should prefer an interpretation which highlights the content of Arg2. This is because the distribution of *but* is biased towards `contra-expectation` with

Arg2 being the semantically salient argument.

- *Although* should rather show less significance as to which of its arguments is more salient. The reason is, when we put *although* in place of *but*, the mid-sentence arrangement of *although* is obtained, which according to the distributional data occurs equally between `contra-expectation` and `expectation` relations. More specifically, we expect that some readers come up with `contra-expectation` interpretation which puts more emphasis on Arg2, and some other readers come up with `expectation` relations which then highlights the content of Arg1.

Another possibility is that the readers do not infer any causal relatedness between the two connected clauses by either of the connectives given the high frequency of both in `Contrast` relations. In this case we expect no effect on the salience of the arguments.

4.4.1 Design and stimuli

In order to see the effect of the connectives on interpretations while keeping the task implicit we design short narrative stories embedding the target discourse relation made by *but* or *although* and ask people to judge the coherence of the entire text. A sample of the stimuli is shown in (32).

(32) **Introduction:** Jane was feeling tired and hungry when she came home yesterday evening.

a. She took some cake from the fridge, **but** she desired to have something savory with her drink.
b. She took some pizza from the fridge, **but** she desired to have something sweet with her drink.
c. She took some cake from the fridge, **although** she desired to have something savory with her drink.
d. She took some pizza from the fridge, **although** she desired to have something sweet with her drink.

Continuation: She had a piece of pizza and went to bed earlier than usual.

Introduction and continuation are kept identical across conditions. Context is changed by alternating cake/pizza and sweet/savory. Each context provides us with a different condition to test people's reaction to the way the story continues. We expect *but* to put an emphasis on the second clause of the discourse relation. Therefore, condition (a) should prepare the reader for accepting the continuation, and condition (b) should result in lower acceptability. According to our hypotheses about *although*, conditions (d) and (c) should be equally acceptable. In other words, context in (a) and (c) is designed to be compatible with the continuation only if a relation with Arg2 emphasis like `contra-expectation` is inferred from it, whereas context in (b) and (d) is only compatible with the continuation if an Arg1 emphasizing relation like `expectation` is inferred.

If no such differences is observed across conditions, then the detailed observations we made with regard to the distribution of *but* vs. *although* across fine-grained relations are not useful for predicting the effect of each connective on semantic processing. In that case, a generalization over different senses of the connective would be a rather minimal and sufficiently explanatory approach.

As a pretest of the stimuli, we included four additional conditions without the continuation sentence to check for the local coherence of the relations. If stories in one condition are incoherent without the last sentence, then the above interpretations would not be valid. Therefore, 24 items with a total of 8 conditions have been designed. Stories are about common situations in which a protagonist has to make a decision. The introduction opens the story with asserting a Question Under Discussion (Roberts, 1996), that is a goal for the protagonist to achieve. In the above example, the protagonist is hungry. The middle part of the story provides more specific information about the situation , e.g., that some cake is available and Jane wants to have something sweet. The continuation sentence provides an ending to the story by telling the final action or decision of the protagonist.

The stimuli is distributed between subjects in a 2 (connective) * 2 (context) * 2 (with/without continuation) design. Each of the 8 lists include 24 actual items and 26 filler items with similar stories to avoid learning effects. 16 fillers use other discourse connectives, shorter or same size stories (2-4 sentences), and coherent/incoherent semantic connections. The remaining 10 fillers are items from another experiment with relatively longer stories (5 clauses). In each list, items were randomized and every participant is given a single list.

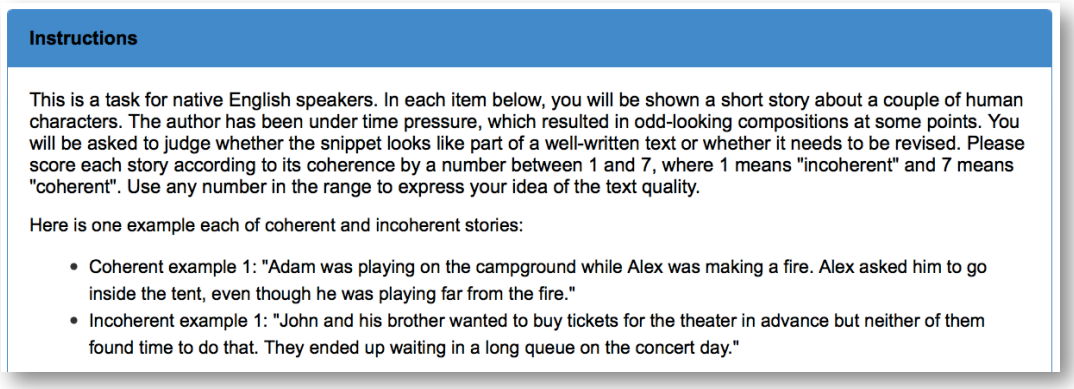


Figure 4.2: Instructions for the Amazon Mechanical Turk users (the coherence judgment task)

4.4.2 Procedure

For this experiment, we recruited 48 native speakers of English on Amazon Mechanical Turk. Each list was published as a HIT within the standard interface of the AMTurk website. Participants were asked to chose a single HIT from the batch, but we also controlled user IDs to double check that no participant took part in the experiment more than once. The instruction of the task is displayed in 4.2. We tried to make an imaginary cover story about the author of the text snippets (i.e., our items) to let the subjects find the source of incoherence themselves, rather than pointing them directly to the suitability of the connective or the continuation. Readers were asked to score each story's coherence in a Likert scale: 7 for coherent to 1 for incoherent. Each story is displayed with a radio box preset to "unanswered".

Among the recruited subjects, 25 were female and 23 were male with average age of 38.25 (min: 22, max: 68). They were all native speakers of English.[7] As compensation, we paid $\$2.5$ per questionnaire. People took between 9 minutes and 3 hours to complete an HIT, as the coherence judgment task was not time-pressured.

4.4.3 Data treatment

In total 2400 samples (48 participant * 50 items) were collected. Only 7 samples were left unanswered in total. Coherence judgment scores for the very incoherent and perfectly coherent filler items were checked to make sure that the participants answered the

[7] Among all, 5 participants indicated that they could speak other languages in addition to English.

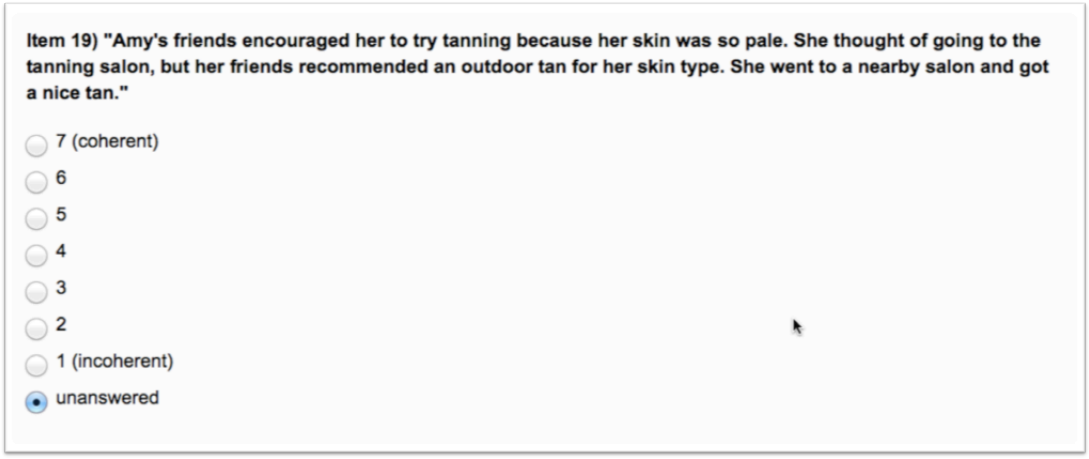


Figure 4.3: An item of the coherence judgment task

questions in a sensible manner. A strict error analysis shows that no participant answered more than 38% fillers unexpectedly, therefore, we did not discard any participant's data. However, some suspicious filler and actual items were double-checked for unintended ambiguity with our native speaker colleague and we changed them slightly for the other two experiments.

4.4.4 Results

Pretest (short versions): We start by explaining the coherence scores obtained for the short version of the stories (excluding the continuation sentence). Coherence scores assigned to the stories containing either of the connectives does not show any significant effect of the connective. A trend is observed toward better average score for *but* sentences as table 4.5 indicates. Also, we looked into the effect of context (the two versions of each story item). Neither Anova nor a mixed-effect linear regression considering participant and item as random effects, and connective and context as fixed effects indicate that the coherence score of the stories made by *but* vs. *although* are significantly different in terms of local coherence. Particular items turned out to have been scored lower by average, which we recognized as indicator of imperfect English, but they were not excluded from the analyses.

Main results (full stories): Table 4.6 presents the average scores obtained for the complete versions of stories from different conditions. Among the *but* conditions, stories with a last sentence confirming an emphasis on the content of Arg1 got a low average coherence score, and stories in which the last sentence attends to the content of Arg2 got a high coherence

Table 4.5: Coherence judgment test (1): scores by connective

Connective	Mean score	SD	Min	Max
although	5.14	1.82	1.00	7.00
but	5.34	1.72	1.00	7.00

score. A one-way anova confirms the main effect of the context on acceptability of the *but* stories ($p - value < 0.001$). On the other hand, no significant effect of the context can be observed in the *although* conditions, yet a trend for higher acceptability again for continuations confirming Arg2 is noticeable.

Table 4.6: Coherence judgment test (1): scores by context and connective

Condition	Mean score	SD	Min	Max
Arg1Emphasis:but	**3.31**	1.80	1.00	7.00
Arg2Emphasis:but	**5.38**	1.71	1.00	7.00
Arg1Emphasis:although	4.52	1.91	1.00	7.00
Arg2Emphasis:although	4.85	1.86	1.00	7.00

By putting together all conditions and fitting a mixed-effect linear model considering participant and item as random effects, similar results are obtained. In this model, a negative main effect of *but* on coherence scores is significant. The effect is to the opposite direction of the trend we found for short version stories, and indicates the importance of the final sentence in determining coherence of the text. Furthermore, a positive interaction between *but* and Arg2 emphasizing context shows up in the complete model, which emphasizes the strong effect of context only on acceptability of the *but* sentences. Throughout the forward model selection procedure adopted for fitting the multi-layer regression model, we tried adding random effects of participant and item separately. It turns out that a considerable variance is captured by introducing random intercepts per participant and per item and the model fit improves according to the ANOVA test of the models. The fixed effects explained above are still highly significant (see Table 4.7) after subtracting the variance caused by individual differences (48 participant) and differences between experimental items (24 stories). Considering participant-specific random slopes reveals a negative interaction between the intercepts and the slopes, which indicates that participants who usually give higher scores to the stories are affected less by the coherence conditions. In this model we had to introduce a variable called *Condition* to capture the four-way variations of the stories, because otherwise the model would not converge if *Connective* and *Context* and their interaction were used for random slopes separately. Again the main effects remain

highly significant (see Table 4.8). In simple words, all statistical analyses run on this data result in the same conclusion that in the *but* conditions, subjects are quite sensitive to the the story continuation emphasizing the content of either of the relational arguments, whereas in case of *although* the coherence score of the entire story is less affected by the way each argument is highlighted.

Table 4.7: Fixed effects in the linear regression model (without random slopes).

LMR fixed effects	Estimate	Std. Error	DF	t-value	
(Intercept)	4.52	0.19	93.64	23.30	***
ContextArg2	0.33	0.19	503.35	1.71	.
Connectivebut	-1.21	0.19	503.53	-6.21	***
ContextArg2:Connectivebut	1.73	0.27	503.35	6.30	***
$lmer(answer \sim Context * Connective + (1\|participant) + (1\|item))$					

Table 4.8: Fixed effects in the linear regression model (with random slopes).

LMR fixed effects	Estimate	Std. Error	DF	t-value	
(Intercept)	4.52	0.25	35.69	18.18	***
ConditionArg1Emphasis:but	-1.21	0.24	42.30	-5.01	***
ConditionArg2Emphasis:although	0.33	0.33	29.53	1.02	
ConditionArg2Emphasis:but	0.85	0.26	25.11	3.27	**
$lmer(answer \sim Condition + (1 + Condition\|participant) + (1 + Condition\|item))$					

4.4.5 Discussion

The pattern we found in this experiment is compatible with the predictions of the distributional account regarding the effect of *but* and *although* on highlighting the content of the relational arguments. In particular, according to the corpus study*but* appeared frequently with a subset of concessive relations that reject Arg1's content by introducing a denial in Arg2. This made us predict that in any context where the two clauses can be put into different relations, *but* should deliver the distributional information associated with it and bias interpretation towards the relation with Arg2 emphasis. This is what we observe in the results of our experiment. Stories continued with an additional sentence pointing to the content of Arg1 are scored as less acceptable compared to their counterparts in which the Arg2 content is being confirmed by the final sentence.

The second hypothesis was, if *but* is replaced with *although*, then the semantic hierarchy should be less effective in prioritizing the message of one argument over the other. This is because the distribution of mid-sentence *although* across relations of different types indicates a balanced proportion as of which argument should be semantically more salient. The expected symmetry showed up in the coherence scores collected from the *although* conditions in our stimuli. As we found, *although* fits well in stories with continuation confirming the content of either arguments. Nevertheless, the average acceptability scores given to the *although* cases is less than the average acceptability of the coherent *but* stories (compare the three bottom rows of Table 4.6). This raises a question of why *although* sentences should be altogether less coherent in the context we experimented. Is it because the connective is not used in a place that perfectly fits its ideal, namely, the most frequent function? This is a question we hope to answer in our second experiment, which includes two conditions with the sentence-initial *although*.

We also observed a general positive trend in acceptability of the context emphasizing Arg2 regardless of the utilized connective. This made us suspicious about a possible confound effect: It could be that the context designed for these conditions when put together with the continuation is more coherent with respect to other semantic criteria such as collocation of specific words, events or concepts. In order to control for this unwanted effect, we use another set of continuation sentences to counterbalance the global coherence effect in our second experiment. Furthermore, we add two more conditions with *but* in which the order of arguments is changed. If stimuli is balanced in terms of the default emphasize on Arg1 and Arg2 we should be able to see exactly the opposite effect for *but* when its arguments are exchanged.

4.5 Experiment 2: different arrangements of *but* and *although*

The following questions need to be answered before we can claim the findings of our previous experiment align with the predictions of the distributional hypotheses:

1) Does *although* result in a lower average coherence because of its non-specificity? That is, whether the low scores in either context is a result of the flat distribution of the connective across `expectation` and `contra-expectation` relations? According to the *but* results, the `contra-expectation` relation is perfectly inferred when *but* is used. When we compare statistics of *but* and mid-sentence *although*, this result aligns very well: *but* is a more specific marker of `contra-expectation` and it should work better

than *although* in a context that supports a `contra-expectation` interpretation. Nevertheless, we have not yet tested the mid-sentence *although* against a baseline connective that perfectly marks the other interpretation, i.e., the `expectation` relation. Therefore, we cannot still argue that the low score for the `expectation` relations is due to the connective and not due to the context itself.

2) Does *but* result in a higher coherence because of its specificity? Talking about the incoherent case made by *but* is easier: since in the corpus less than one percent of *but* occurrences are with relations with emphasis on Arg1, using this connective in a context where Arg1 is highlighted in the continuation would results in low coherence scores. However, it is not easy to argue yet whether the specific information content of *but* gives rise to one interpretation over the other. Because it could also be that the type of narration emphasizing on the Arg2 content is more coherent than its counterpart emphasizing Arg1. We are in particular suspicious about this issue because of the marginal main effect of the context in our previous experiment: a positive effect ($p - value < 0.1$) was observed for ContextArg2 condition regardless of the utilized connective.

A second experiment is conducted, in order to answer the above question, as well as for looking into the effect of *although* in the sentence-initial position. This experiment provides us with a bigger picture of the way distributional information affects interpretation of the relations that are made possible by *but* and *although*.

4.5.1 Design and stimuli

Stimuli in this experiment are very similar to the one used before. Since we already pretested the coherence of the short version stories, here this condition is excluded. The four **baseline conditions** are same as our previous complete story versions: *but* and mid-sentence *although*.

(33) **Baseline conditions**
Introduction: Jane was feeling tired and hungry when she came home yesterday evening.

a. She took some cake from the fridge, **but** she desired to have something savory with her drink.
b. She took some pizza from the fridge, **but** she desired to have something sweet with her drink.

c. She took some cake from the fridge, **although** she desired to have something savory with her drink.
d. She took some pizza from the fridge, **although** she desired to have something sweet with her drink.

Continuation: She had a piece of cake and went to bed earlier than usual.

Here, we have new conditions with *although* at the beginning, as well as, conditions including *but* with its arguments reversed. Note that the objective of adding each new set of conditions is different: sentence-initial *although* conditions do not change the underlying inference "$P \rightarrow Q$", they are designed to be compared against their *but* and mid-sentence *although* equivalents. Therefore, for making these conditions, we only move the subordinate clause together with its connective to the beginning of the sentence. Therefore, Arg1 and Arg2 are kept the same across all conditions so far. On the other hand, the *but* reversed conditions are made by exchanging the arguments: Arg1 of this set of conditions is Arg2 of all other sets of conditions. This hypothetically should change the underlying "$P \rightarrow Q$", therefore, should result in an interpretation to the opposite of the interpretation of baseline *but* conditions.

(34) **New conditions**

Introduction: Jane was feeling tired and hungry when she came home yesterday evening.

a. **Although** she desired to have something savory with her drink, she took some cake from the fridge.
b. **Although** she desired to have something sweet with her drink, she took some pizza from the fridge.
c. She desired to have something savory with her drink, **but** she took some cake from the fridge.
d. She desired to have something sweet with her drink, **but** she took some pizza from the fridge.

Continuation: She had a piece of cake and went to bed earlier than usual.

In addition to manipulating the design by changes applied to the middle part of the stories, in this experiment, we also changed the continuation sentence of all conditions. This is in particular to resolve our suspicion regarding the main effect of the context

observed in the previous experiment (remember question 2). The new continuations are counterparts of the previous ones. For example, rather than finishing with *"She had a piece of pizza and went to bed earlier than usual."*, in this run, the story is completed with *"She had a piece of cake and went to bed earlier than usual."*, that gives us a fully counter-balanced data when put together with that of the previous experiment. Finally, a total of 8 lists were examined between-subjects in a 4 (connective:arrangement) * 2 (context) design. Similar to the previous experiment, each list contained 24 actual items plus 26 fillers, which were the enhanced version of our previous fillers.

4.5.2 Procedure

In this experiment, we recruited another 48 native speakers of English on Amazon Mechanical Turk. The same instruction was shown to the users and we explicitly asked not to take the test if they submitted a similar HIT before. We also looked into the user IDs to double-check if people adhered to this rule.

Among the recruited subjects, 27 were female and 20 were male with average age of 38.17 (min: 23, max: 68). They were all native speakers of English.[8] As compensation, we paid $\$2.0$ per questionnaire. Like in the previous experiment, people took between 9 minutes and 3 hours to complete a HIT.

4.5.3 Data treatment

In total 2400 samples (48 participant * 50 items) were collected. Only 9 samples were left unanswered. Coherence judgment scores for the very incoherent and perfectly coherent filler items were controlled to make sure that the participants answered the questions in a sensible manner. A strict error analysis shows that no participant answered more than 25% fillers unexpectedly, therefore, we did not remove any participant's data.

4.5.4 Results

Replication of the baseline effects: Table 4.9 displays the results with stimuli including *but* and *although* in the middle and the new arrangement of *although*, all conditions with the new continuation sentence. All effects from our previous experiment are replicated.

[8] Among all, 3 participants indicated that they could speak other languages besides English.

Mid-sentence *but* is judged to be significantly more coherent in the context compatible with a `contra-expectation` interpretation, that is the condition where the continuation emphasizes the content of Arg2. In mid-sentence *although* conditions no significant difference is observed regarding compatibility with a `contra-expectation` or an `expectation` relation like before.

Table 4.9: Coherence judgment test (2): scores by context and connective setup

Condition	Mean score	SD	Min	Max
Arg1Emphasis:but	3.64	1.96	1.00	7.00
Arg2Emphasis:but	5.97	1.54	1.00	7.00
Arg1Emphasis:although	4.95	1.82	1.00	7.00
Arg2Emphasis:although	5.17	1.90	1.00	7.00
Arg1Emphasis:although-initial	6.05	1.36	1.00	7.00
Arg2Emphasis:although-initial	3.49	2.05	1.00	7.00

Specificity of the connective: Comparing the initial *although* conditions with the other two conditions reveals that our previous findings were valid to be interpreted as a result of the connectives' graded specificity. As we expected from the distribution of relations occurring with the initial *although*, this connective fits perfectly in context confirming an *expectation* interpretation, in which the content of Arg1 is highlighted. The high coherence score of this condition removes our suspicion regarding other possible confounds, e.g., that *although* sentences generally make a story more complicated to process, thus less coherent. On the other hand, we see that the Arg2 emphasizing condition is no more acceptable in presence of this connective in the new arrangement.[9] Comparing mid-sentence *although* conditions with the sentence-initial *although* conditions under a linear mixed-effect model (with participant and item as random effects) confirms a significant effect of the argument arrangement on the coherence score (see Table 4.10). People like sentences made by the initial *although* in general better than those made by the medial *although*. The analysis also reveals an interaction between context and arrangement, that is *although* in a mid-sentence position has a bias towards emphasizing Arg2 which is reversed by using the same connective in the sentence initial position.

Fitting a model for *but* and sentence-initial *although* conditions reveals a strong interaction between connective and context (see Table 4.11). A negative main effect for *but* and a negative main effect for the Arg2 emphasizing continuation is observed. However, when the two are combined (the ContextArg2Emphasis:ConnectiveBut condition), the

[9]Remember that Arg2 is always the argument syntactically attached to or introduced by the connective.

Table 4.10: Fixed effects in the linear regression model: Mid-sentence and initial *although* conditions

LMR fixed effects	Estimate	Std. Error	DF	t-value	
(Intercept)	4.95	0.20	89.43	24.21	***
ContextArg2Emphasis	0.22	0.18	498.56	1.21	
Arrangementreversed	1.10	0.18	498.56	6.00	***
ContextArg2Emphasis:Arran.reversed	-2.78	0.26	498.61	-10.74	***
$lmer(answer \sim Context * Arrangement + (1\|participant) + (1\|item))$					

coherence score goes up. This shows that either connective can generate an interpretation that is completely biased towards a specific relation. Like in the previous experiment we adopted a forward model selection procedure and kept the most effective factors in the regression. Adding random slopes did not downgrade the significance of the fixed effects, thus we only reported the results when random intercepts for participant and item have been considered. However, similar to what we observed in the previous experiment, we found a negative interaction between the intercepts and participant-specific slopes which again indicates that optimistic subjects tend to be less sensitive to the manipulations, i.e., coherence of the stories.

Table 4.11: Fixed effects in the linear regression model: Initial *although* and *but* conditions

LMR fixed effects	Estimate	Std. Error	DF	t-value	
(Intercept)	3.64	0.19	101.64	18.91	***
ContextArg2Emphasis	2.33	0.18	502.01	13.14	***
Arrangementreversed	2.41	0.18	502.01	13.57	***
ContextArg2Emphasis:Arran.reversed	-4.88	0.25	502.14	-19.42	***
$lmer(answer \sim Context * Arrangement + (1\|participant) + (1\|item))$					

Symmetry of the context: since the stimuli is designed in a way that both direction causal inference will be possible, we expect to see reversed effect of *but* on highlighting content of either of the connected clauses when the order is reversed. Looking into the coherence scores for baseline and reversed *but* conditions confirms that the context is fairly acceptable for both direction inferences. Table 4.12 shows that *but* in all condition highlights the argument attached to it, that is the Arg2, regardless of what content it has. Therefore, we conclude that the stimuli is balanced and does not include any confound factor giving rise to one over the other interpretation. This result also further elaborates on the key effect of *but* in its context. Given that the context makes it possible to draw a causal inference,

distributional information attached to *but* guides the way each clause's content should be put into which slot of the "$P \rightarrow Q$".

Table 4.12: Coherence judgment test (2): scores for *but* conditions

Condition	Mean score	SD	Min	Max
Arg1Emphasis:but:original	3.64	1.96	1.00	7.00
Arg1Emphasis:but:reversed	3.53	1.98	1.00	7.00
Arg2Emphasis:but:original	5.97	1.54	1.00	7.00
Arg2Emphasis:but:reversed	5.73	1.59	1.00	7.00

4.5.5 Discussion

As we expected, sentence initial use of *although* biases interpretation towards an `expectation` relation between the two clauses. The argument demonstrative of Q' regarding the underlying inference rule "$P \rightarrow Q$" is the more salient statement in this type of relation. We found that stories continued with a sentence confirming P rather than Q' obtained lower coherence scores. The other arrangement of *although*, however, is coherent with either type of continuation (confirming Arg1 or Arg2). This is expected when we consider the distribution of this connective across different types of relations. The mid-sentence *although* is used to same extents in relations that either highlight Arg1 or Arg2. Its lower overall frequency and low specificity to either type of relation can account for the lower coherence scores obtained for this connective in either context (compared with its sentence-initial counterpart in `expectation` and compared with *but* in `contra-expectation` relations).

Our results on *but* rejects the under-specified account trying to assign a general role to this connective that covers all its usages. In the concessive context that we designed, we do not see an effect of *but* that can be called a simple contrast effect, if we assume that this term refers to the most general relation sense that *but* marks covering all negative polarity relations. Given the symmetry assumption, simple contrast should not prefer the content of one clause over another, whereas *but* does: People disliked stories including *but* when the last sentence was compatible with an *expectation* interpretation.

We also controlled for the effect of the context sentences. The context supports a causal interpretation, however, it does not dictate a specific direction. Even if it does, the effect is easily overridden by the utilized connective. Using *but* in the middle of the two clauses has the same influence on the interpretation regardless of the clausal content,

i.e., *but* always puts emphasis on the content of Arg2, which takes the role of Q' in the "$P \rightarrow Q$" inference.

Findings of both experiments suggest that specificity of a connective to a relation in terms of their co-occurrence frequency can account for the way the connective biases interpretations of a text towards inferring that particular relation. Not only the tendency for inferring a certain relation is determined by the connective specificity (*but* for `contra-expectation` vs. *although* for `expectation`), but also the amplitude of the effect (bigger tendency of initial *although* for `expectation` compared to that of mid-sentence *although*) can be determined based on the frequency data.

4.6 Experiment 3: online effect of *but* vs. *although*

The two experiments we studied so far in this chapter show that interpretation of readers from connected sentences in short stories are affected by the fine-grained meaning properties of the utilized discourse connectives. From previous work we know that connectives prepare readers for immediately upcoming context (the relational argument attached to the connective). Our experiments additionally proved that using even very closely related connective types such as *but* and *although* can make specific context sentences (i.e., the story continuation) more or less compatible, and thus affects the global coherence of the text. The relevance of the effects to online comprehension processes is still a question. In this section, we investigate online relational processing to see whether readers pick up on the difference between *but* and *although* during reading or whether it becomes clear to them only when they are explicitly asked to judge a text's coherence. If human readers are sensitive to the different properties of *but* and *although* as reflected in their distribution across discourse relations, we should find that:

- When *but* is used in a relation, a continuation emphasizing the content of Arg2 should be read smoothly and faster than a continuation emphasizing Arg1, as the distribution of *but* and the results of the offline experiment suggest.
- When *although* is used exactly in the same relation, the two types of continuation should be read more similarly than if *but* is used. This is because neither the distribution of the mid-sentence *although* nor our coherence judgment test revealed a big preference of this connective as to which of its arguments should be highlighted.

4.6.1 Design and stimuli

The stimuli in this experiment is the same as what we used in experiment 1 excluding the short story conditions (ones without the continuation sentence). But we test each item also with the alternative continuation sentences designed for experiment 2 to have a fully counterbalanced design. This gives us a total of 8 conditions: 2 (connective) * 2 (context) * 2 (continuation). Here is an example of the items in different conditions:

(35) **Introduction:** Jane was feeling tired and hungry when she came home yesterday evening.

a. She took some cake from the fridge, **but** she desired to have something savory with her drink.
b. She took some pizzda from the fridge, **but** she desired to have something sweet with her drink.
c. She took some cake from the fridge, **although** she desired to have something savory with her drink.
d. She took some pizza from the fridge, **although** she desired to have something sweet with her drink.

Continuation 1: She had a piece of cake and went to bed earlier than usual.
Continuation 2: She had a piece of pizza and went to bed earlier than usual.

The two factors context and continuation will be collapsed in the analysis, because continuation 1 has the same function for conditions (b) and (d) as continuation 2 has for conditions (a) and (c), and having both continuations only aims at obtaining a counterbalanced design.

An eye-tracking-while-reading experiment is designed to examine online reading behavior of the subjects in exposure to 24 stories like the above example. Each participant sees each item only in one condition (that means 3 items from each of the 8 conditions). Items are mixed with filler stories, as well as items of two other experiments with almost similar number of sentences and narrative content. In total every participant reads 84 stories (including 24 *but/although* items, 12 fillers and 48 other experimental items) and after each story answers a YES/NO comprehension question. For example, the question designed for the above item is as follows:

(36) Was Jane at home the entire yesterday?

Questions designed for the 24 *but/although* items are all about the introduction part of the stories, therefore the right answer does not depend on any inference based upon the discourse relation or the continuation segments. Each participant receives equal numbers of questions with YES vs. NO answers. Fillers had questions from all different parts of the text to avoid directing the listeners where to fish the answers throughout the experiment. Comprehension questions have two purposes. First, to evaluate the number of right answers to make sure a subject has been conscious and engaged in reading; second, to analyze both the reaction time and the correctness of the answers to see if they are correlated in any way with the difficulty of the passages, i.e. the coherence conditions. Otherwise, our main focus is on the reading patterns we collect by tracking the eyes of subjects while reading the stories.

In addition to the total reading time of a story and reaction to the question, reading time measures for specific chunks of the story are collected separately. These specific areas include the connective area, the variable areas containing the critical content in Arg1 and Arg2 of the relation, and the final sentence. In the following example, all critical areas are highlighted.

(37) Jane was feeling tired and hungry when she came home yesterday evening. She took some **cake** from the fridge, **but** she desired to have something **savory** with her drink. She had a piece of **cake** and went to bed earlier than usual.

Some items have critical areas with longer phrases. In eye-tracking-while-reading experiments, it is desired that the critical areas across all items are as similar as possible. It means, while the content is changing, all critical areas that will be compared later in the analysis should have moderately similar length and should be viewed on the same location of the screen, ideally on the same line. Furthermore, critical areas should not fall near the end or beginning of a line, as otherwise they might be skipped in natural reading. Designing 24 items complying with these criteria is not trivial and needs a lot of edition and correction. For example, we had to modify our items from the offline experiments by removing and adding filler words and phrases and get revisions from a native English speaker to make sure the sentences were modified coherently. The result is a set of aligned items with almost similar length and position of critical areas, but to some extent different

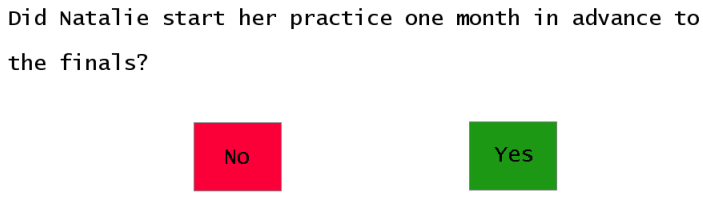


Figure 4.4: Screen shot of an example comprehension question

from the original sentences.

4.6.2 Procedure

The eye-tracking experiment is implemented within the Experiment Builder software for an Eye-link 1000 Plus tracker. The experiment consists of two main blocks of item presentation: practice and actual trial. Practice items are similar to the actual experimental items but eye-tracking data is not collected for them.

Each block starts with an instruction screen telling the subject how to read the stories, how to proceed to the comprehension questions and how to answer them. All text material on the screen are viewed in `Lucida Console` font (with same length characters), size 20 and triple line spacing. Figure 4.4 views an screen shot of a comprehension question. Subjects are asked to press the space key after reading a story to navigate to the question screen, and press J and F keys for YES and NO answers, respectively. On the question screen readers are given a clue of left and right to find the place of the key without having to look at the keyboard (to avoid head movement).

A total of 39 native English speaker subjects are recruited for the experiment at the University of Edinburgh. Subject were invited to the eye-tracking lab and asked to sign a consent form before taking part in the experiment. We paid each subject 12 pounds in compensation for a maximum of two hours (including the eye-tracking experiment and two other experiments). They were seated in front of a display computer equipped with a desktop mounted camera and a head rest. The experimenter sited in front of the eye-tracker host to control the experiment. Subjects were asked to put their chin on the rest, press their forehead to the upper part, and try not to move their head during the experiment. Camera calibration was performed once at the beginning of the experiment and whenever the experimenter noticed big drift during the experiment. A drift check screen with a dot in the middle was displayed before each story screen to give the experimenter the

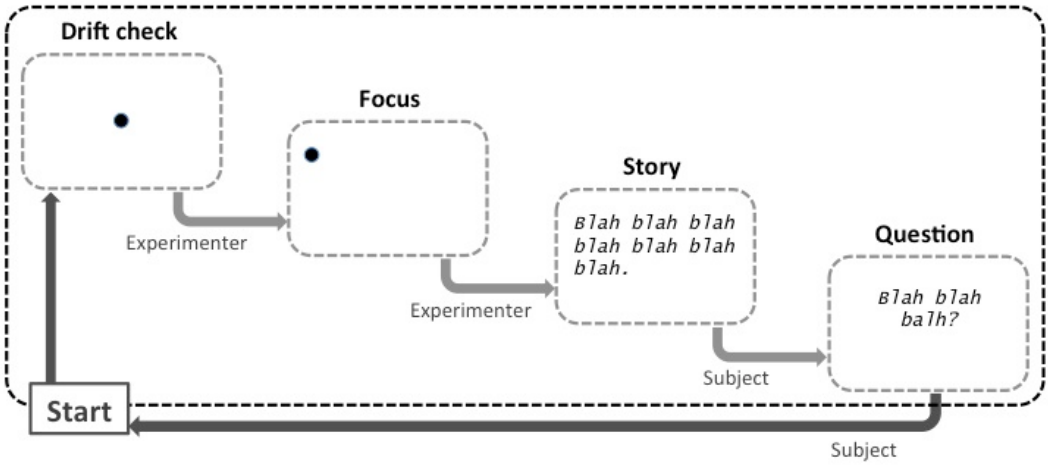


Figure 4.5: Sequence of screens viewed to the subjects in the eye-tracking experiment

choice to pause eye-tracking and re-calibrate.[10] Figure 4.5 shows a simplified schema of the experiment event sequence including the drift check screen and a focus screen where people were asked to look at a point appearing at the beginning of the first line of the story. The experimenter controlled these two transitions from the host computer. Then the story was displayed and the control for moving to the question screen is handed to the display computer, i.e., the subject. By pressing the answer key (F or J) on the question screen, the entire sequence started for the next experimental item.

The eye-tracker general settings are summarized in Table 4.13. We track both eyes when possible but use data from the dominant eye for our analysis. Data collected by the eye-tracker include the coordination and duration of every fixation on the screen plus saccades, i.e. the rapid movement of eye between fixation points, as well as short duration track losses detected as blinks.

The eye-tracking experiment is followed by a memory test, to measure the subject's memory span. This is a standard type of test that our colleagues designed for us.[11] The purpose is to measure how subjects keep text content in their memories. The memory span can play a confound role in the correctness of answers to the comprehension questions and also might affect reading patterns, e.g., when a sentence contradicts with a non-immediate but related sentence in the preceding part of the text. A third experiment is also included in the session which is not relevant to our study, but rather to the filler experiment. In the

[10] Also, in case the experimenter decided to give the subject a break the experiment is paused on this screen (depending on the duration it took the participant to read the stories, they were given one or 2 breaks in the middle of the eye-tracking experiment).

[11] Thanks to Merel Scholman for designing the stimuli and the web-based framework.

Table 4.13: Eye-tracker general settings

Property	Velue
Tracker version	Eyelink 1000 Plus
Camera mount	Desktop
Mount usage	Binoc/Monocular - Stabilized head
Eye event data	Gaze (Fixations, Saccades, Blinks)
Eye tracking mode	Pupil - Corneal
Sampling rate	1000 (500 each eye) per Second
Calibration	9-sample model

following section we will talk about the data collected in the eye-tracking and the memory test. All these experiments together took between one and two hours depending mostly on the subjects reading pace, and difficulty of camera setup for them.

4.6.3 Data treatment

We had to discard data from a few subjects because of frequent head movement, blinks or longer track losses during the experiment (which happens due to mascara, dark lashes, or anti-reflex glasses). Reading between lines also makes it difficult for us to assign fixations to specific interest areas, thus a few subjects were also removed in the data preprocessing phase. Preprocessing of the data before we run statistical analyses includes the following steps:

- Drift correction: in this experiment the camera is mounted by the desktop rather than on a helmet. Therefore, it happens often that when the subject moves his head, the collected data points are drifted from their actual coordinations. This happens even if we perform several camera calibration procedures during the experiment. A manual drift correction is thus required after data collection, specially for fixing the y coordinations in a reading experiment (where data from one line of the text might be mixed with data from another line due to vertical head movement). I used the Data Viewer software from the SR Research products for vertical drift correction. Every trial of every participant is viewed on the screen as a video with a map of interest areas overlaying it. If some fixation is located in a wrong region — i.e., based on watching the eye-movements they belong to some other region — it should be manually moved. In our experiment no horizontal drift correction is applied.

- Fixation removal and merging: eye fixations have different durations and due to

blinks and track losses they might fall in the wrong places. Before data analysis we need to make sure that the fixations we consider for calculating reading time measures are part of the reading process. I removed fixations larger that 800 and combined fixations smaller than 80 ms with the closest neighboring fixation in one-character distance.

Data from 32 out of 37 subjects remained after filtering and corrections.

4.6.4 Results

This section presents the results of our online reading experiment. First, we look at the coarse-grained results on the total reading time of the stories and comprehension questions. Second, we present an analysis of the correctness of the answers given to the comprehension questions; at the end, a more detailed analysis of reading behavior on the specific interest areas of the stories is performed.

Overall reading times: Stories took between 6.7 and 37.5 seconds to be read over all conditions and participants with an average of 16.6 seconds per item. We have a range of slow to fast readers, as well as a range of less to more time consuming items. In all per-condition analyses in this section we will use mixed effect regression modeling in which participant and item are considered as two random effect factors. The reading time of the stories are compared across conditions in Figure 4.6. No significant difference can be observed between conditions in terms of the total time spent for reading the stories. The same applies to the reading time of the questions. Questions took between 0.7 and 12.1 seconds, and a mean of 2.8 seconds to be read and answered, but no significant difference is observed between conditions, see Figure 4.7. This suggests that the coherence of the stories in terms of the compatibility of the final sentence with the relation constructed by the connective does not influence reading times at this coarse scale. The memory span size of the subject was also included in both models for sentence and question RT, however it also did not show up as an effective factor.

Question answering correctness: as aimed by design, questions varied in terms of difficulty and it shows up in the average correctness of the subjects answers. Two questions were removed from this analysis because we discovered that in one of them the name of the protagonist was not mentioned correctly, and in the other we found a temporal ambiguity resulting in lower than 50% correctness across subjects. All other questions obtained higher than 50% correct answers (ranging between 52% and 100%). Participants

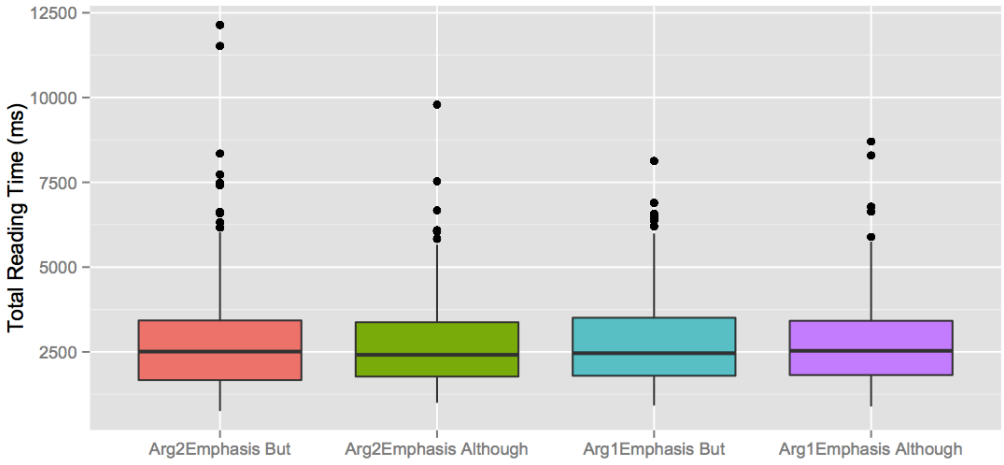


Figure 4.6: Total reading time of stories across four conditions

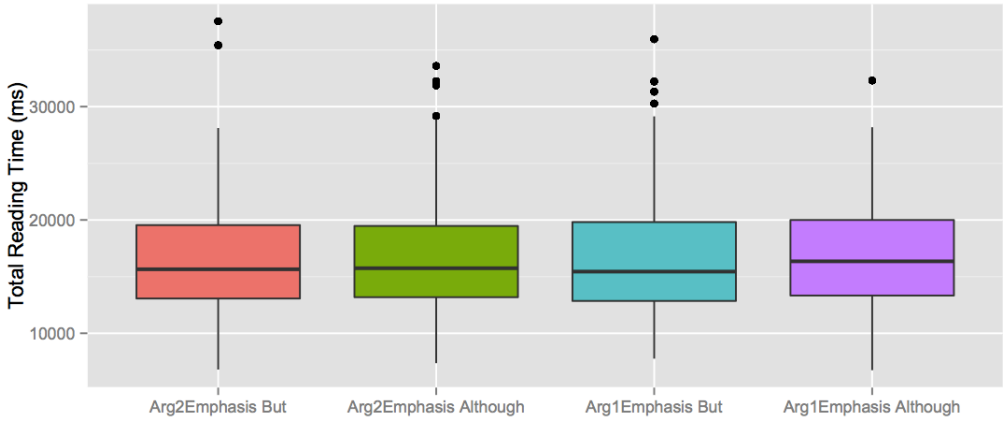


Figure 4.7: Total reading time of questions across four conditions

Table 4.14: Answers to the comprehension questions

Condition	Correct answer (mean)	Correct answer (sd.)
Arg1Emphasis But	**0.79**	0.41
Arg1Emphasis Although	0.81	0.40
Arg2Emphasis But	**0.85**	0.36
Arg2Emphasis Although	0.83	0.38

Table 4.15: Fixed effects in the linear regression model of answer correctness (*but* conditions).

LMR fixed effects	Estimate	Std. Error	DF	t-value	
(Intercept)	0.41	0.12	33.28	3.43	**
ContextArg2Emphasis	0.06	0.04	303.90	1.68	.
MemorySpanSize	0.08	0.02	27.48	3.32	**
$lmer(Correctness \sim Context + MemorySpanSize + (1\|participant) + (1\|item))$					

show various levels of performance in answering the comprehension questions. They score from 50% to 100% correct answers with mean and median of 82% (most people answered more than 18 questions correctly). We observe an effect of subjects' memory span size on their performance in answering comprehension questions ($p - value < 0.05$). Total reading time of the story and the question do not correlate with correctness of the answers. On the other hand, coherence conditions do. Based on the results of the coherence judgment study, emphasis on the content of Arg1 should be more interrupting in the *but* condition, whereas *although* should be unbiased with respect to which of its arguments will be emphasized in the final sentence of the story. Table 4.5 shows the proportion of correct answers to the comprehension questions across conditions. Correct answers are most likely when the *but* relation is followed by Arg2 emphasizing context and least likely when it is followed by Arg1 emphasizing continuation. A similar but smaller difference is observed between the corresponding *although* conditions. The effect of context is marginally significant (Arg1 emphasizing context leads to less accurate answers), and connective does not show up as a significant factor or interaction. However, fitting a mixed effect regression with all factors (connective, context and participant's memory span) as fixed effects plus participant and item as random effects only reveals a significant main effect of the effect of memory span size on the correctness of the answers. People with larger memory span provide more correct answers to the questions. The best fit obtained through a forward model selection procedure reveals only a marginal effect of the story coherence for the subset of data including *but* (see Table 4.15). In other words, the varience introduced by participant-specific factors is high and after accounting for it in the model the other fixed effects do not show up as important. Including any other interaction between the fixed and random factors does not improves the model fit. As a summary, this analysis shows that answers to the comprehension questions (which are about the earlier part of the story and independent of the variable inferences) are only influenced when the text is obviously incoherent (connective *but* in Arg1Emphasising context) and the reader has a small memory span size.

Table 4.16: Eye-tracking experiment: final region reading time

Condition	Total duration (mean)	Total duration (sd.)
Arg1Emphasis But	**332.92**	266.05
Arg1Emphasis Although	328.48	263.06
Arg2Emphasis But	**296.28**	235.46
Arg2Emphasis Although	331.90	346.72

Interest areas: We did not see any effect of the coherence on the total reading time of the stories, which is not surprising for an online comprehension study. Effects in eye movement data are usually very small and local, therefore an investigation of the critical areas of the text is required. Based on our hypothesis regarding the inferences triggered by the connectives and results of the coherence judgment test, we expect that the reading behavior should start to vary between conditions as soon as the critical area in the final sentence of the story is encountered ("She had a piece of **cake** and went to bed..."). Connectives can also have local effects on processing of the second argument of the relation but this effect is not the main interest of the current study. I start by analyzing the eye-movement measurements collected on the critical area of the final sentence, e.g., on *cake* in the following example:

(38) Jane was feeling tired and hungry when she came home yesterday evening. She took some **pizza/cake** from the fridge, **but/although** she desired to have something **savory** with her drink. She had a piece of **cake** and went to bed earlier than usual.

Several eye movement measures are used in reading studies within specified interest areas: total duration (sum of all fixations in a given area), go pass time, also known as regression path duration (sum of all fixations in the area and regressions to the previous areas before moving to a following area), first pass duration (sum of all fixation in the area the first time it is visited until it is exited to any other area), and finally the number of regressions in or out of a given region that are independent of the duration of fixations but are not very frequent eye movements in smooth reading. A regression-out is counted when the interest area is exited to a preceding area (to the left in English) before a following area is fixated. A regression-in is counted when the interest area is entered from a following area. The main measure in our analysis is the total duration, while interesting patterns observed in other measures will be mentioned too.

The critical area in the final part of the story includes one to four words across items. When averaged over items we can look into the differences in milliseconds spent on reading of this area and approximate its processing difficulty. Table 4.16 compares the total duration of the critical area in the final sentence of the story across coherence conditions. We find that the most time consuming condition is the incoherent one according to our previous experiments. In fact, patterns of processing difficulty in terms of reading time are as we expected: *but* emphasizes on Arg2, therefore an Arg2 emphasizing continuation is easier and an Arg1 emphasizing continuation is more difficult to process, whereas for *although* both contexts are almost equally acceptable. Interestingly, these differences are visible on the earliest word of the final sentence where a semantic reference is made to the content of Arg1 and Arg2 of the preceding relation, that is, as soon as the word *cake* in the above example elaborates what the protagonist finally decided to do.

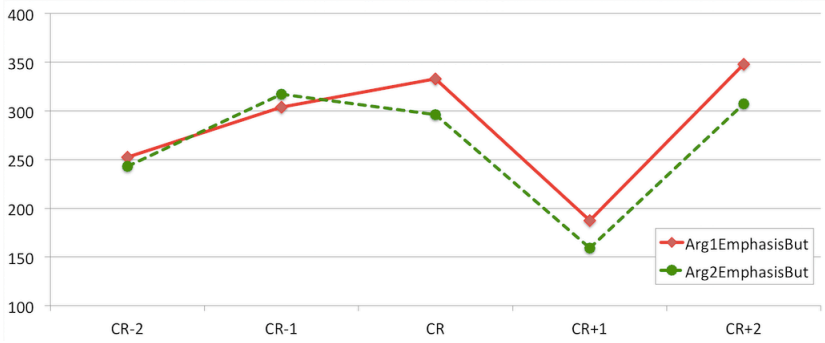


Figure 4.8: Total reading time of the critical region and its neighboring regions in the final sentence.

Figure 4.8, depicts the total reading duration of this critical area and its neighboring areas in the *but* conditions. While in areas previous to the critical phrase average reading times are almost equal across conditions, encountering the critical content results in a deviation in the reading time of the rest of the story, as it can be seen in the graph. Difference between the two *but* conditions is marginally significant at the critical area ($p < 0.1$). Also, regressions out of this area occur significantly more often in the Arg1 emphasizing condition than the Arg2 emphasizing condition ($p < 0.05$). Interestingly, the two *but* conditions also differ in terms of the total reading duration ($p < 0.1$) and regression-in ($p < 0.05$) of the area right before the critical phrase in Arg2, i.e., *sweet/savory* in the above example. Increased regression-in is also observed at the onset of Arg1 for the less coherent *but* condition ($p < 0.1$). Regressions from the final sentence to the previous context, i.e., the relational arguments, indicate that the subject re-read that part of the story. This provides more evidence for the Arg1 emphasizing continuation to be an incoherent, thus difficult to process context for relations made by *but*. All these effects are discovered

Table 4.17: Eye-tracking experiment: final region regression-out

LMR fixed effects	Estimate	Std..Error	df	t.value	
(Intercept)	0.21	0.04	72.83	4.90	***
ConnectiveBut	0.02	0.05	610.15	0.33	
ContextArg2Emphasis	0.04	0.05	612.79	0.84	
ConnectiveBut:ContextArg2Emphasis	-0.14	0.07	613.38	-2.04	*
$lmer(RegOutCount \sim Connective * Context + (1\|participant) + (1\|item))$					

by fitting mixed effect linear regression models that consider participant and item as random factors. While in *but* conditions we observe a significant difference of context on reading time measures, such a clear pattern does not show up in data from the *although* conditions. The interaction between connective and context factors in a mixed-effect model fitted to the entire data (on regression-out of the final CR as the dependent variable) indicates that *but* strongly prefers a continuation that confirms its Arg2 and *although* is rather non-selective. No main effect of the context is observed on the investigated measures (see the fixed effects in the full regression model in Table 4.16).

4.6.5 Discussion

The eye-tracking experiment revealed an effect of the coherence relation introduced by the two connectives *but* and *although* on reading time of critical regions of the text and correctness of the answers to the comprehension questions. In particular, we found that in *but* stories, continuations emphasizing on the content of the first argument are more difficult to process and more distracting in question answering. This is what we expected based on the corpus-based modeling of the meaning of *but* and the results of our coherence judgment study. However, the effect size is small in all our measurements, so a strong conclusion about the way distributional properties of a connective can be effective in online processing of short narrative text is not possible. We did not find any difference in processing of the stories with different continuations when *although* was used. This is not surprising, as the extreme cases were the *but* conditions which I just talked about. All these weak results, when compared against the very strong effects we found in the offline study, suggest that in online reading people might be less sensitive to the global coherence factors than when they are asked explicitly to rate the coherence of a text. More online studies need to be conducted though, to make sure that the findings of our eye-tracking experiment is representative of what is happening in people's mind during online processing of text.

In other words, eye-tracking might not be the ideal methodology to investigate this level of sentence processing or our weak result might be due to insufficient number of subjects.

4.7 Summary

Results of the experiments in this chapter confirm the hypotheses we posit based on the distribution of *but* and *although* across relations in PDTB. In particular, we find that in a context where both connectives fit locally, each triggers to a different interpretation that is the more frequent fine-grained relation the connective co-occurs with. The finding rejects an under-specified account of connective meaning, one that assigns a core meaning to a multi-sense connective like *but* and leaves the specification to be determined by the content of the arguments (Fraser, 1999). We see that the connective affects the salience of the involved clauses through the inference of a specific relation. In using *although*, the argument immediately following the connective (the syntactically attached clause) is given less importance as a result of its occurrence in discourse enforcing this pattern. On the other hand, *but* emphasizes its following clause more since it has a bias towards other relations. Interestingly, when *although* is used in the middle of the two arguments, its function converges to that of *but*, which is predicted based on the distribution of the medial *although* across relations of different types in the corpus. Running lab experiments in parallel with a corpus study reveals a neat probabilistic pattern indicative of the relation between comprehension and production data. Not only the general bias of each connective in prioritizing clause-level content reflects the distribution of the connective across discourse relations of different types, but also more detailed distributional properties (degree of specificity of the connective to the relations it often marks) and the arrangement of the arguments (difference in the distribution of medial vs. initial although) turn out to determine the size of biases we find in our coherence judgment experiments. The results we obtained from this offline study are very clear and show the significant effect of connective usage profile on under-specified context (one that allows different interpretations). The online study reveals weaker tendencies. In particular, we expected to see processing difficulty or cost of expectation update to be proportional to the information delivered by the connective about the discourse relation congruent with the context, but we only found a very small difference between same context being processed when *but* vs. *although* were used. This could be due to the high sensitivity of reading time measures to noise and confound factors, or due to the relatively small subject population in the online reading study (32 subjects were tested in the eye-tracking experiment vs. 48 in each of the offline coherence judgment

experiments).

This is the first comparative study of two connectives in English with closely related meanings that uses natural production data to predict effects on comprehension and examines these predictions experimentally. The objective of our study is more general, though. Examination of *but* and *although* is carried out, in the first place, to elaborate on the information theoretic account of discourse cues. In particular, we proved that the distributional representation of connective meaning is helpful to determine similarities and differences between multi-sense connectives in a unified and quantified framework, which is missing in theoretical pragmatics. Generalization and specification in a distributional representation can be made by changing the granularity of the dimensions, i.e., discourse relation senses used for annotation of the reference data source. In PDTB three levels of granularity have been provided; other discourse relation categorization systems can also be adapted as long as annotated data is available for them. Otherwise, unsupervised machine learning approaches need to be applied to discover the type of relation that the connective co-occurs with. Current automatic methods heavily rely on the connective itself to determine the discourse relation between two sentences (see the review of implicit relation identification systems in Chapter 2). Therefore, we still can't reliably represent the meaning of a discourse connective based on automatically extracted relations/dimensions. One of the directions for the future work is to find a feature space based on words and linguistic features in the context that represents discourse cue meaning, similarly to what is being done for modeling content word meaning in distributional semantics. This will be a shift from the dictionary-like tradition of defining the meaning or the type of contexts where a discourse cue is used, to an information theoretic perspective that gives each cue type a probability of occurrence in any given discourse context. Discourse relations will then be the hidden states of the model analogous to the underlying inferences humans are engaged with during discourse comprehension. In such a computational model, the linguistic devices traditionally regarded as discourse cues would be part of the surface representation of inferences just like other words and phrases. In other words, every expression can be viewed as a contributor to relational inference and thus can be regarded as having a relational meaning representation.

Chapter 5

Discourse connectives modulate information density

In the previous chapter, we showed that the natural distribution of discourse connectives in different contexts, namely in various discourse relations, can give us a quantified answer to how differently two connectives in the same context bias the reader's interpretation of the story. This chapter focuses on a second question: in which circumstances does a speaker use discourse connectives to make relations explicit and in which not?

According to our general analysis of PDTB (Section 3.2.4), about half of the annotated relations for neighboring sentences in this corpus do not include any explicit discourse connective, despite the fact that the annotators discovered these relations and even found suitable connectives for them. Furthermore, relations of different types show significantly different proportions of explicit and implicit occurrences, thus explicit marking of a relation does not seem to be a random decision. This is an unexplored phenomenon that we will try to explain via an information theoretic approach. The *Uniform Information Density* theory — along with other theories on efficient communication introduced in Chapter 3 — proposes that optional markers in a language should be left out naturally if they would lead to a trough in information density, and be inserted in order to avoid peaks in information density (Levy and Jaeger, 2007). In this chapter, we investigate whether this principle can explain patterns of discourse connective utilization and omission in natural text. The effective factors that make a discourse relation more or less predictable are first identified. The information that a discourse connective adds to its context needs to be measured with respect to the predictability of the relation. Previous studies on default **interpretation biases** and predictability of relations in certain **linguistic contexts** are used for this purpose. A set of experiments is conducted on finding the correlation between each of these two factors and connective reduction. We find that some relations, which according to previous theories are expected by default, tend to appear more often without explicit connectives as opposed to generally unexpected relations. Also, predictability of a relation given the linguistic context, i.e., strong features in the first discourse segment of a relation, turns out to play a role in the writer's choice to drop discourse connectives.

5.1 Relation predictability and linguistic marking

The tradition in the information theoretic approach to language is to formulate information content based on predictability: the less predictable events (symbols, words, structures) are the more informative ones. In order to investigate discourse relations from an information theoretic perspective, we first need to identify the main factors that affect the predictability of a discourse relation. The most obvious factor is language-internal:

- The **linguistic encoding** of the discourse relation. Semantic and structural features that are extracted from the two discourse segments or their larger context, plus overt discourse connectives can make a relation sense more likely to be inferred than others. See Chapter 2, where we introduce these linguistic features in the context of automatic discourse relation identification.

In addition to the linguistic encoding of a relation, some language-external factors might also affect the predictability of a certain relation in a given context. These include:

- The specific **world knowledge** of the listener about the events being narrated. For example, the listener might have read a story and know how specific events were related to one another.
- The **cognitive biases** that affect language comprehenders while interpreting relations between consecutive sentences in a text or an utterance. Some relations might be preferred by default over others because of ease of inference or general expectations.

Speakers do not have any control over the latter two factors, but they can make a discourse relation easy to infer by means of linguistic encoding. Thinking of the reader-side comprehension of the discourse relation as a Bayesian inference, we can model the language-external factors as shaping the prior expectation of a relation, and the linguistic marking by the speaker as the factor determining the posterior probability of a relation sense to be inferred:

$$\text{predictability of a relation} \propto \text{linguistic marking} * \text{prior expectations}$$

Mathematically speaking:

$$p(r|markers) = \frac{p(markers|r)}{p(markers)} * p(r) \tag{5.1}$$

Modulating the information density at the level of discourse relations can be viewed as changing the probability of the relations by manipulating their linguistic markedness. This can mainly be done by using suitable discourse connectives. Other markers are subject to grammatical and semantic constraints involved with construction of clausal units. Thus the choice of form in production of a discourse relation can be simplified in the following way: speakers can either use an optional discourse connective to specify the relation between two discourse segments explicitly, or leave the relation implicit for the reader to be inferred by the help of non-optional linguistic features in the context and their prior expectations.

If the communication principles that we introduced in Chapter 3 regarding efficient delivery of information between speakers and listeners are in function at the level of discourse relations they should affect the way discourse relations are expressed. We attempt to examine the discourse connective utilization patterns within the framework of the Uniform Information Density theory. As we pointed out previously, one of the predictions of the UID theory is that the optional markers in a language should be reduced when the structures they signal are predictable (Levy and Jaeger, 2007). Thus the hypothesis we are going to test is the following:

> Assuming that the speakers have a listener model and formulate their utterances in the way that is both efficient to produce and easy to comprehend, they should drop discourse connectives in predictive context. On the other hand, when the relation is not predictable, discourse connectives should be used to avoid unexpectedness and processing difficulty at the listener side. In summary, a rational speaker model dictates **the less predictable a relation is the stronger marking or explicitation would be required**.

The alternative hypothesis would be that speakers do not consider comprehender-side capacities while producing a linguistic signal, or that the discourse production is not guided by such comprehension-driven procedures. If this is the case, no correlation would be expected between the predictability of a relational structure and the way explicit marking is performed. Looking into naturally uttered sentences or naturally generated text helps us understand what the real strategy of language producers are.

The above formula 5.1 also defines a basis for measuring the linguistic markedness of a relation r in which a specific discourse connective is utilized:

$$markedness(r) \propto \frac{p(C|R)}{p(C)} = \frac{p(C,R)}{p(C) * p(R)} \tag{5.2}$$

where R is the intended relation to be inferred and C is the connective. These probabilities can be calculated based on the distribution of the discourse connectives and discourse relations in a reference corpus. In this chapter, a large-scale study is conducted on PDTB. Regarding the way we distinguished among factors affecting identification of discourse relations, experiments are organized in two parts. First, Section 5.2 looks into previous psycholinguistic work on the cognitive biases that are effective during text comprehension. *Continuity and causality-by-default* hypotheses (Segal et al., 1991; Murray, 1997; Levinson, 2000; Sanders, 2005) propose that readers have a default bias for causal interpretations and linear deictic shift. We explain these theories in detail, and based on the redundancy account propose that causal and continuous relations should appear more often without explicit connectives since readers expect them by default, whereas non-causal and discontinuous relations which are unexpected given the prior should be made explicit by using the connectives. This hypothesis is then examined by analyzing the implicit and explicit relations in PDTB, as well as a more focused analysis of markedness degree of the explicit relations. Second, in Section 5.4, we focus on the predictability of the relations given the linguistic features. Considering that the comprehenders receive the relational arguments incrementally, we examine the non-optional linguistic predictors of the discourse relations in their first clause, i.e., before the connective is encountered. The question is whether or not the presence of such markers correlates with reduction of the optional connectives. Two sets of linguistic features are examined with respect to previous work: Implicit Causality verbs (as markers of the`reason` relations) and negation words (as markers of the `alternative` relations). These relations together with their cues are extracted from PDTB to test the UID-based hypothesis regarding the omission of discourse connectives.

5.2 The effect of cognitive biases

This section looks into the predictability of a relation based on the general expectations readers incorporate while reading a text, as to in what order and in what way events should be narrated and related to one another. We will first give an overview of the theories

and cognitive research on interpretation biases and then investigate whether patterns of explicit marking in the PDTB corpus of discourse relations can be explained based on the predictability of relations according to the reviewed theories. In order to do so, We propose a way of measuring the information content of discourse cues and use it for a unified and quantified corpus analysis.

5.2.1 The continuity hypothesis

Based on a large body of studies on narrative understanding and presumptive meaning, Segal et al. (1991) proposes a principle of continuity according to which a new sentence is interpreted to be continuous to its preceding context unless discontinuity is explicitly marked.

Continuity and discontinuity in this theory are defined with regard to the notion of *deictic shift* (Bruder et al., 1986): *"within the world of the story, the reader may be required to shift from one established deictic center that is a certain time, place or character focus to another deictic center."* A set of linguistic markers including discourse connectives are identified to signal discontinuity, meaning that they provide readers with information that the frame of referents needs to be shifted. Segal et al. (1991) look into the preference of adult readers when they interpret relations between consecutive sentences in short narrative texts. They find some initial evidence for continuity preference in that subjects tended to infer additive relations between sentences in the stories. Additive relations are pure examples of continuity according to the theory. Comparing causal connectives *so* (as a temporally continuous) and *because* (as a temporally discontinuous) in a connective placement experiment, they find that people used *so*, three times more often than they used its backward counterpart (in the original texts there existed 10 *so* and 5 *because* cases). The tendency to infer continuous relations is influenced when original connectives are shown to the subjects. That is, people managed to detect discontinuous relations in places where the relation would not be detected if the connective was not present.

Murray (1995; 1997) tries to explore the default bias for continuous relations by looking into the comprehension difficulty involved with construction of the discontinuous relations. Reading experiments conducted by Murray (1995) reveal a greater facilitating effect of adversative connectives compared to causal or additive connectives, each in its own suitable context. In a second set of experiments, Murray (1997) tested the inappropriately placed adversative, causal and additive connectives. Relation of each type is examined with the markers of the other two relations as well as with no marker. In all relations, the

no connective condition is read faster, thus this condition is considered as the baseline. Compared to causal and additive relations, adversative relations cause more reading disruption when a wrong or no connective is used in them. Also, the most difficult condition among causal and additive relations are the ones including an addversative connective. In other words, inappropriately placed adversative connectives caused more RT disruption than either inappropriately placed additive or causal connectives. Murray (1997) interprets this as an evidence for the continuity. This is a controversial argument, given that it could only be the similarity between additive and causal relations that make them easier material to process when their connectives are used interchangeably. An offline sentence production experiment in the same study provides better evidence for the continuity hypothesis. In this experiment, a sentence is given to the subjects and they are asked to provide a continuation. There are four conditions, one for each type of connective (additive, adversative and causal) plus a no-connective condition. Murray (1997) finds that in connective-present cases people provide continuations consistant with the relation the connective marks. In the no-connective conditions, most of the continuations are causal, followed by additives and finally very few adversative relations. Murray attributes the difference between the additive/causal connectives and the adversative ones in all above mentioned experiments to the underlying discontinuity of the discourse relations the latter group marks. According to this theory, causal and additive relations are more expected by default when no explicit relational marker is present in text, and adversative connectives are more salient in text because they override the default expectation of the reader for continuous relations.

Levinson (2000) indicates that when events are narrated one after another in text, they tend to be read as temporally successive and if plausible, as causally linked. Regarding temporal relations, the continuity hypothesis predicts that shift in the time frame should be considered as discontinuity. Therefore, a simple additive relation is more continuous than a temporal relation. But among different ways of expressing events occurring in different times, the forward temporal relations should be more expected than the backward relations. Thus, cues for temporal non-linearity such as *after*, as opposed to *before*, which indicates the expected temporal order, should be more salient given that they mark a backward temporal shift. Murray classifies causal relations as continuous ones, but he also notes that a connective like *because*, which signals a temporally non-linear causal relation (*backward* transition from the effect to the cause), should have stronger contextual effects than connectives such as *so* or therefore for similar reasons. Concessive relations that are marked by connectives like *although* and *however* are considered as *negative causal* relations (König, 1991). According to Segal et al. (1991) and Murray (1997), such

relations should not be expected to the same degree as positive causal relations, which benefit from a higher degree of continuity.

5.2.2 The causality-by-default hypothesis

The causality-by-default hypothesis by Sanders (2005) is a stronger argument regarding the expectedness of causal relations: "because experienced readers aim at building the most informative representation, they start out assuming the relation between two consecutive sentences is a causal relation". A special status is given to causal relations in terms of how informative they are compared to other ways sentences can be inter-connected in a listeners mental representation of the text. The review of comprehension studies in the previous chapter provides some indication regarding the importance given to causal relations and causal discourse markers in psycholinguistics. Older studies on narrative comprehension view story understanding as an attempt by the comprehender to find cause-effect relations among the narrated events (Trabasso et al., 1982, 1984; Trabasso and Sperry, 1985). That is why dead-end events in a story are less important compared to other events that have some cause or consequence narrated in the context (see the experiment in Trabasso and Sperry (1985)).

The effect of local causality on processing of consecutive sentences in a text has been investigated in several experimental studies. Murray (1997) found a default preference for causal relations in a sentence completion task where subjects were asked to continue a given sentence in the way they wanted. While in connective-present conditions (1-a), continuations were consistent with the relation the connective signaled, in the connective-absent condition (1-b), subjects tended to provide continuations that stood in a causal relation with the given sentence.

(1) a. Ronny cleaned up the house for his girlfriend's visit, so/and/but ...
b. Ronny cleaned up the house for his girlfriend's visit. ...

Keenan et al. (1984) conducted a reading experiment with stimuli composed of two sentences. The first sentence varies across conditions between four levels of causal relatedness to the second sentence. They found that the stronger the causal relation between the sentences, the shorter it took people to read the second sentence. A recent EEG reading study by Kuperberg et al. (2011) came up with similar results. Small discourse

consisting of three sentences was easier to process when the sentences were causally related. Specifically, a larger N400 (an EEG signal which typically indicates semantic anomalies) was found when sentences were irrelevant. All of these findings suggest that readers have a prior expectation that consecutive sentences in a text should be causally related and congruent. However, the default expectation can be altered if explicit cues such as a concessive connective provide marking for other types of relations (Drenhaus et al., 2014; Köhne and Demberg, 2013; Xiang and Kuperberg, 2014; Xu et al., 2015).

Some arguments against the causality-by-default hypothesis can also be found in the literature. Millis et al. (1995) performed an experiment where two consecutive sentences (that did not stand in an obvious causal relationship) were connected with a full stop, or one of the three discourse connectives *because*, *and* or *after*, as the indicators of causal, additive and temporal relations, respectively. The sentence pairs inherently could be interpreted as expressing any of the mentioned relation types. Millis et al. found that causal inferences (as measured by asking participants a "Why?" question after pair of sentences) were only reliably made in the *because* condition, but not in the conditions where the sentences were connected by a period or one of the other connectives. They concluded that the discourse marker *because* played a very important role in people's forming of an inference, and that this inference was not formed automatically contrary to the prediction of the causality-by-default hypothesis.

Findings of the reviewed experiments are dependent on materials. For example, how likely in a relation the second sentence is to be a cause or consequence of the first sentence by content, which may differ between studies and explain contradictory results. A study of natural production data that is annotated systematically with discourse relations may help us compare different relation senses with respect to their frequency and degree of implicitness. Given that a corpus is annotated by a certain groups of people and according to a fixed annotation schema, a large-scale study of this type will hopefully resolve some controversies shaped based on scattered laboratory findings.

5.3 Experiment 1: connective reduction in causal and continuous relations

The studies on continuity and causality that we reviewed are all small scale and use carefully designed experimental materials. It is an open question whether the hypotheses regarding the default bias for constructing continuous and causal relations between consecutive sentences also hold for naturally occurring texts. The corpus-based experiment in

this section investigates the validity of a prediction made based on these theories when put together with the general idea of connective reduction: if continuous and causal relations are expected by default they should often appear without explicit marking. All other things kept equal, a discourse marker would be a more redundant cue in a predictable relation than in other contexts.

Cognitive bias or default expectation for specific discourse relations could be interpreted in different ways. Specific relation senses might be actively looked up by language comprehenders as a step towards achieving a coherent mental representation of the text, or specific relations might be inferred because they are less expensive than others to be processed (in terms of consuming certain cognitive resources such as memory). In either case, we predict that the speakers should tend to reduce the optional markers of the expected relations more often than they reduce the optional markers of the other relation types. This means that we expect a higher ratio of explicit occurrences or average markedness (that will be defined later) for the less expected relation types compared to the more expected ones.

5.3.1 Data selection

For this experiment, we look into all explicit and implicit relations in the PDTB corpus. As we explained in Chapter 2, annotators of PDTB have used the same set of relation sense tags to label both explicit and implicit relations in the corpus. Explicit relations are identified by the help of discourse connectives in the original text (2-a). About 5% of these relations (999/18459) have been annotated with two relation tags, indicating that the connective could convey two relations at the same time (2-b).

(2) a. There have been no orders for the Cray-3 so far, *though* the company says it is talking with several prospects.
– `Comparison.Contrast`

b. *When* the fruit is ripe, it falls from the tree by itself.
– `TEMPORAL.Synchrony & CONTINGECY.Condition.general`

Implicit relations, on the other hand, are discovered by examining whether two neighboring but unconnected sentences in the text could be put into a discourse relation by using an artificially inserted discourse connective. These connectives are annotated in PDTB (3).

Among 16053 implicit relations, 171 are annotated with two alternative connectives (3-b). For the purpose of annotation, each inserted connective is then treated just like original connectives in explicit relations, namely, one or two relation sense(s) from the tag set are chosen as its label(s). Among the inserted implicit connectives 360 instances are tagged with two relation sense labels.

(3) a. The government counts money as it is spent; [whereas] Dodge counts contracts when they are awarded.
– `Comparison.Contrast`

b. Regulators are wary. [For example/because] They haven't forgotten the leap in share prices last Dec. 7.
– `EXPANSION.Restatement.specification`
– `CONTINGENCY.Cause.Reason`

In order to compare relations of different types regarding their reduction of connectives, we first look into the ratio of explicit to total occurrences of each relation sense (Section 5.3.4) and then a weighted average over the information content of the connectives used for marking instances of the relation (Section 5.3.5).

5.3.2 Mapping from PDTB to continuity and causality

In this section, we propose some heuristics for classification of PDTB relations into continuous, discontinuous, causal and non-causal. Categorization of relations in PDTB is different but we find a mapping between that and the causality/continuity space by using a set of primitive features of discourse relations (Sanders et al., 1992). These feature include the **basic operation** of the relation (causal/additive), the temporal or logical **order** of the events in the two arguments (basic/ reversed) and the **polarity** of the relation (negative/positive).[1] Recall that in PDTB, relations are categorized in three levels of granularity: classes (level 1), types (level 2), sub-types (level 3). I explain the types and subtypes wherever they differ from the parent class regarding a primitive feature values. All relations that we discuss in this section are displayed in Table 5.1 with the values of

[1] Sanders et al. (1992) additionally count the **source of coherence** in a relation (semantic/pragmatic) as a primitive feature All PDTB relations except the ones named as *pragmatic*, like the `Pragmatic cause` are semantic relations by definition. For now, we assume that the pragmatic version of a relation obtains the same set of other attributes as that of its semantic version.

their primitive features.

The **basic operation** distinguishes between causal relations and relations in which the two arguments cannot be put into any causal relatedness, i.e., the additive ones. The PDTB definition of the relation class CONTINGENCY and one of the types categorized under COMPARISON, i.e., Concession, indicate that these relation senses together with their finer grained sub-types should be attributed with a *causal* basic operation:

> The class level tag CONTINGENCY is used when the connective indicates that one of the situations described in Arg1 and Arg2 causally influences the other.
>
> The type Concession applies when the connective indicates that one of the arguments describes a situation A which causes C, while the other asserts or implies C'.

All other branches in the hierarchy should be assigned the value *additive* for the basic operation dimension. In particular, all EXPANSION and TEMPORAL relations, as well as Contrast from the COMPARISON class are additive because they do not involve any causal inference:

> The class EXPANSION covers those relations which expand the discourse and move its narrative or exposition forward.
>
> The class TEMPORAL is used when the connective indicates that the situations described in the arguments are related temporally.
>
> The type Contrast applies when the connective indicates that Arg1 and Arg2 share a predicate or property and a difference is highlighted with respect to the values assigned to the shared property. In the Contrast relation, neither argument describes a situation that is asserted on the basis of the other one.

In Sanders et al. (1992), **order** is only defined for relations that have a causal basic operation and it pertains to the way the causal inference is made. The order is basic if the information in the first discourse segment (Arg1) expresses P, and nonbasic if the first discourse segment expresses Q in the underlying $P \rightarrow Q$ operation. Sanders et al. (1992) indicate that additive relations are symmetric, therefore, order of the segments does not discriminate between different types of additive relations. we argue that even if the basic operation is additive, a relation can be assigned a certain order with respect to the temporarily of the events being narrated. In PDTB, a separate class is defined for TEMPORAL relations. The sub-types in this class are defined on the basis of the temporal order in which events in the two arguments are sorted. In succession Arg1 talks about

Table 5.1: Primitive features of the PDTB relation senses

PDTB relation	Basic operation (Causality dimension)	Order (Continuity dimensions)	Polarity
COMPARISON.Contrast	additive	none	negative
COMPARISON.Concession.expectation	**causal**	**reversed**	**negative**
COMPARISON.Concession.contra-exp.	**causal**	**basic**	**negative**
CONTINGENCY.Cause.reason	**causal**	**reversed**	**positive**
CONTINGENCY.Cause.result	**causal**	**basic**	**positive**
CONTINGENCY.Condition	causal	basic	positive
EXPANSION.Conjunction	additive	none	positive
EXPANSION.Instantiation	additive	none	positive
EXPANSION.Restatement	additive	none	positive
EXPANSION.Alternative	additive	none	**negative**
EXPANSION.Exception	additive	none	**negative**
EXPANSION.List	additive	none	positive
TEMPORAL.Asynchronous.precedence	**additive**	**basic**	**positive**
TEMPORAL.Asynchronous.succession	**additive**	**reversed**	**positive**
TEMPORAL.Synchronous	additive	none	positive

an event happening before the one narrated in Arg2, and in `precedence` the relation is the other way around. [2] Therefore, we can directly use the definition of these relations to identify the value for the order attribute in our mapping.

> The tag `Synchronous` applies when the connective indicates that the situations described in Arg1 and Arg2 overlap.
>
> As sub-types of `Asynchrony`, `precedence` is used when the situation in Arg1 precedes the situation described in Arg2, and `succession` is used otherwise.

The distinction made between the two sub-types of `Asynchrony` is very similar to the distinctions made between sub-types of `Cause` and also sub-types of `Concession`. Each of these relation types covers two sub-types that only differ in the order dimension (all highlighted in Table 5.1). The sub-types `result` and `contra-expectation` correspond to forward or basic order that we see in `precedence`, whereas `reason` and `expectation` correspond to backward or reversed order which relates them to

[2] Order of a relation is also discussed extensively in (Knott, 1996) with several meanings. Here we only consider the logical order of events in causal relations and the temporal order. Whenever the first applies, we ignore the second. In causal relations, these two orders are usually the same. Exceptions can be exemplified, e.g., for pragmatic causal relations: "Mary did not show up in the party, because she will have an exam tomorrow".

succession:

> The type reason is used when the situation described in Arg2 is the cause and the situation described in Arg1 is the effect, and result applies when the situation in Arg2 is the effect brought about by the situation described in Arg1.
>
> Two Concession sub-types are defined in terms of the argument creating an expectation and the one denying it. Specifically, when Arg2 creates an expectation that Arg1 denies, it is tagged as expectation. When Arg1 creates an expectation that Arg2 denies, it is tagged as contra-expectation.

Finally, the **polarity** dimension determines the way the truth of an argument typically determines the truth of the other. A relation is negative if one argument is involved in a causal or additive basic operation with the negative counterpart of the other argument. The definition of COMPARISON suggests that all of its types should be considered as negative polarity relations. Among other senses in the hierarchy, two types of EXPANSION, i.e., Exception and Alternative, encode a sense of negative polarity:

> The class tag COMPARISON applies when the connective indicates that a discourse relation is established between Arg1 and Arg2 in order to highlight prominent differences between the two situations.
>
> The type Exception applies when the connective indicates that Arg2 specifies an exception to the generalization specified by Arg1.
>
> The type Alternative applies when the connective indicates that its two arguments denote alternative situations.

I consider other relations in the hierarchy as positive polarity relations. This applies to all senses under CONTINGENCY and TEMPORAL, as well as, List, Restatement, Conjunction and Instantiation among the second-level senses, and finally, conjunction and disjunction among the third-level sub-types of EXPANSION. Table 5.1 summarizes the feature assignment to the PDTB relations. Some relations fall into the same bucket if we only consider a feature spaces including these three dimensions. Specially, EXPANSION relations in which order does not take a value, all look alike. It is important to note that these relations are different with regard to other criteria which might affect their degree of continuity. In particular, Restatement is the most symmetric and neutral relation in the corpus:

> In `Restatement` the semantics of Arg2 restates the semantics of Arg1. It is inferred that the situations described in Arg1 and Arg2 hold true at the same time (*in fact, in other words*).

On the other hand, `List` and `Instantiation` enforce some progress or shift in perspective:

> The Type `List` applies when Arg1 and Arg2 are members of a list, defined in the prior discourse.

> The tag `Instantiation` is used when the connective indicates that Arg1 evokes a set and Arg2 describes it in further detail.

Making a decision about the `Conjunction` relation is not easy given its vague definition. This label sounds to have been given as the last choice of the annotator:

> The Type `Conjunction` is used when the connective indicates that the situation described in Arg2 provides additional, discourse new information that is related to the situation described in Arg1, but is not related to Arg1 in any of the ways described for other types of `EXPANSION`.

`Conjunction` covers a huge proportion of explicit and implicit relations in the entire corpus (about 25%) and looking into its instances does not help us find a certain dimension for distinguishing it from the parent category, thus not much can be said about its degree of continuity.

5.3.3 Predictions

Now let's get back to our study of causal and continuous relations. Relations that obtain the *causal* value in the **basic operation** column of the table should be classified as causal relations. If these relations are expected by default, connectives should often be removed in these relations to avoid redundancy. While concession relations are taken as causal types, they indicate violation of defeasible causal relations between events in the real world. Therefore, not only compared to their positive polarity causal counterparts, but in comparison to additive relations they should be less expected. Therefore, we predict the following patterns of connective omission in causal relations:

- `Cause` relations (including both sub-types `reason` and `result`) should tend to appear in natural text without explicit discourse connectives because they are expected by default.

- `Concession` relations (including both sub-types), should tend to occur with explicit marking because they are among the most unexpected relations that a reader could infer or they are totally the opposite of the default expectations.

`Condition` is also a class of causal relations but in English can only be implicated with very specific syntactic constructions. The explicit marker of this relations (*if*) cannot be dropped in the way most discourse connectives can. Therefore, we do not include `Condition` in our connective omission analysis.

Classification of the relations into continuous vs. discontinuous can be performed by examining the two features **order** and **polarity**. We take negative polarity of a relation as a sign of discontinuity in discourse. In negative polarity relations, Arg2 asserts an statement that is unexpected, atypical, or complementary to the content of Arg1. The other type of discontinuity is related to the temporal order of the events being narrated. In both basic and reversed order relations, a degree of discontinuity exists because of the time shift between the two arguments, but reversed relations introduce even a higher degree of discontinuity due to violating the default occurrence order of events. Predictions that result from this categorization are as follows:

- Among `EXPANSION` relations `Exception` and `Alternative` should tend to appear in natural text with explicit marking because they are discontinuous given their negative polarity. Other types, including `Restatement`, `List` and `Instantiation` are continuous and should appear more often without connectives if continuity is expected by default. No specific prediction can be made for `Conjunction` relations.
- All relations from the `COMPARISON` class should tend to occur with explicit connectives because they are discontinuous and therefore unexpected. `Concession` should be less predictible and therefore more explicit compared with `Contrast`. Among `Concession` relations, `expectation` should be the one with highest rate of connective usage, given that it reverses the temporal relation between the two arguments besides having a negative polarity.
- `TEMPORAL` relations should occur more often with connectives because they encode shift in time. Among them `succession` should benefit most from explicit marking given that it encodes a reversed order of events.
- Among `CONTINGENCY` relations, `result` is the one that keeps the expected continuous order of events, therefore, it should need less explicit marking than its backward counterpart, `reason`.

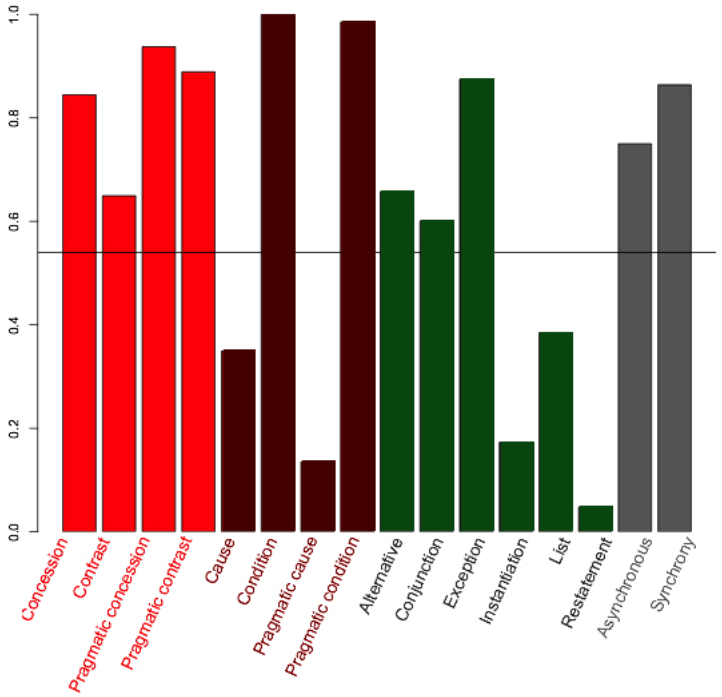


Figure 5.1: Connective use ratio in level-2 PDTB relations (% of explicit cases per relation)

In some explicit relations Arg2 appears first in the text as an outcome of a fronted use of the discourse connective. This is because the annotators were asked to annotate the syntactically attached arguments to the connective as Arg2, and the other one as Arg1. For example, *because* and *although* can take their subordinate clause (Arg2) to the beginning of a composite sentence:

(4) ***Because*** [ARG2: *the drought reduced U.S. stockpiles*], [ARG1: *they have more than enough storage space for their new crop, and that permits them to wait for prices to rise*].

These relations need a special treatment when we analyze temporal continuity. A `reason` relation with a fronted connective should be considered as continuous and a `result` relation of this type should be considered as discontinuous. The same applies to the concessive and temporal equivalents of these relations.

5.3.4 Connective use ratio analysis

To investigate the above hypotheses, the first the easiest method would be to compare the number of times each relation in PDTB occurs with and without discourse connectives. We define the ratio of connective use for a given relation r as follows:

$$\text{Connective use ratio (r)} = \frac{\text{Frequency of explicit } r}{\text{Total frequency of } r}$$

Explicit relations in the corpus that are annotated with two sense labels (5% of the explicit relations in the corpus) are counted for each label independently. For example, a sentence including *while* annotated by `TEMPORAL.Synchrony` and `COMPARISON.Contrast` is considered in calculation of the above measure for both relations. Implicit relations with two sense labels (2% of the implicit relations in the corpus) are also treated in the same way.[3]

Figure 5.1 compares the ratio of connective use across second-level relations types in PDTB. Relations belonging to the same high-level class are colored the same. A line indicates the weighted average, that is the ratio of all explicit relations to the total number of explicit and implicit relations in the corpus (54%). A taller bar means that the relation often appears in text with an explicit discourse connective and is indicative of the unexpectedness of the relations. A shorter bar, on the other hand, means that in most occurrences of the relation, connectives have been left out by the writer which based on our hypothesis should happen when the relation is expected by default. It is time to see whether or not hypotheses regarding default expectation for causal and continuous relations are consistent with the patterns of connective omission in the corpus:.

Causality hypothesis: The causality-by-default hypothesis (Levinson, 2000; Sanders, 2005) proposes that people prefer to interpret consecutive sentences as standing in a causal relationship. If that is true, causal relations should not need explicit marking and writers should drop the connectives to express these relations more often than in other relations. As we see in the Figure 5.1, `Cause` is not the only relation that is often implicit, and hence with view on the reduction hypothesis, it sounds like the causal relations are not the only predictable relations. `Instantiation` and `Restatement` from the `EXPANSION` class tend to be constructed without connectives too. Nevertheless, `Cause` and `Pragmatic cause` together constitute the most frequent type of implicit discourse relations in the PDTB. The connective use ratio of causal discourse relations (0.35) is significantly lower than the that of other frequent discourse relations, in particular `Conjunction` (0.60), `Contrast` (0.65), `Asynchronous` (0.75), as well as the average overall relation types (0.54), – all comparisons significant according to a

[3]For 171 implicit relations that constitute only 1% of all implicit relations in the corpus two candidate discourse connectives are annotated and for each separate relation senses are assigned. In these cases, we consider each of these connectives separately and count them the same we count explicit relations with one or two annotated relation(s).

binomial test with $p - value < 0.001$. On the other hand, the negative polarity causal relations, i.e., Concession which violate a default expectation of causality between two events strongly prefer to be expressed explicitly, i.e., with connectives (0.84). This result is consistent with our predictions based on the causality-by-default and the connective reduction theories: connectives tend to be reduced more often in causal relations to avoid redundancy because readers would be expecting this type of relation by default.

Continuity hypothesis (polarity): In the previous section we classified relations in PDTB with respect to continuity or discontinuity of an event by looking into two cognitive primitives, polarity and order of the operation underlying the definition of the discourse relations. In particular, we proposed that all of the discourse relations in the COMPARISON and TEMPORAL family should be considered as discontinuous. Within the EXPANSION family, we classified Instantiation, Restatement and List to encode a higher degree of continuity than Exception and Alternative. Comparing these relations in Figure 5.1 reveals a strong correlation between the connective use ratio of a relation and its continuity classification. All COMPARISON and TEMPORAL have higher rate of explicit marking (avg. 0.79 and 0.69, respectively) than the average of all relations. This pattern applies to Exception (0.88) and Alternative (0.66) too, whereas other EXPANSION types exhibit a much lower rate of connective use compared to average: Instantiation (0.17), Restatement (0.4) and List (0.38). The PDTB data thus provides strong supporting evidence for the continuity hypothesis combined with the idea of connective reduction as a means of avoiding redundancy.

Continuity hypothesis (temporal order): Now it is time to compare the subtypes of Cause, Asynchrony and Concession relations to investigate whether continuity in the temporal ordering of events is implicit or marked explicitly. As we mentioned earlier, we need to take care of the order of the arguments in this analysis, because some explicit relations have fronted connectives. In case of implicit relations, annotators have always inserted a connective in between the two arguments. Table 5.2 presents frequency information on the Arg1-connective-Arg2 versus the connective-Arg2-Arg1 occurrences of explicit Cause, Concession and Asynchrony. Interestingly, there are always more ordinal modifications (the connective-initial presentation) when a temporally backward relation of any type is being expressed. This implies that even in the presence of the cues, people have a tendency to keep the textual order of the arguments the same as the temporal order in which the associated events happened. This observation might also be indicative of a general force behind natural convergence of grammar in English: gradually shaped syntactic rules that let the application of *because* and *although* at the beginning of a com-

posite sentence might be an influence of cognitive preferences, as they ease comprehension of the clauses by ordering them in a linear way. In case of *although*, the sentence-initial formulation is even dominant and more constraining in terms of the inference it produces (see the previous chapter's experiments on different usages of *although*, Section 4.5).

Type:subtype (explicit only)	Arg1-Conn-Arg2	Conn-Arg2-Arg1	Signif.
Cause:result	746	6	***
Cause:reason	1324	**164**	
Concession:contra-expectation	791	13	***
Concession:expectation	183	**209**	
Asynchrony:precedence	931	55	***
Asynchrony:succession	867	**234**	

***: significant pairwise differences according to a binomial test (at $p < 0.001$)

Table 5.2: Distribution of textually ordered vs. reversed occurrences of arguments in causal, concessive and temporal relations with explicit connectives.

In order to conduct an accurate analysis of the temporal transition given the information about argument organization, we combine all implicit and explicit occurrences of connective-initial and medial instances from all the 6 relations and perform a correlation analysis. In this analysis connective-initial occurrences of backward relations count as temporally continuous and connective-initial occurrences of forward relations are taken as temporally discontinuous (e.g., a `reason` relation in which Arg2 appears first in the text is taken as continuous, just like a `result` relation in the form of Arg1-connective-Arg2). A chi-square test of relations' temporal continuity and their ratio of connective use shows a significant correlation between these two factors ($\chi^2 = 67.31$, $df = 1$, $p < 0.001$). This observation supports our hypothesis based on the continuity and connective reduction theories: temporal continuity is expected by default by the listener when they processes consecutive sentences, therefore its explicit cues are more likely to be dropped by speakers. On the other hand, temporally discontinuous relations, which are not expected by default, need more explicit marking.

5.3.5 Markedness analysis

The analysis in the previous section only looks into the presence and absence of the discourse connectives. As we saw in Chapter 3, connectives of different types have different information contents in terms of how frequently they occur in specific relations

or how distinctively they help the readers to distinguish one relation from the others. Therefore, the mere presence of *a* discourse connective might not be a perfect indicator of that relation instance to be explicitly marked. A good example is the connective *and* that is used in a variety of explicit relations. This connective, despite of being present, does not mark a relation as strongly as a connective like *because* does. Here, we try to define a scalar measure of *markedness* for a relation instance to overcome this drawback of the binary approach. The terms unmarked and marked are used in linguistics to distinguish between a regular, simpler, more common or easier to produce form of a linguistic construct from a more specific form. Hume (2004) argues that markedness has been always a vague notion, and elaborates that in fact predictability is the basis of markedness and motivates a quantified rather than a descriptive approach to markedness. This applies to the case of discourse relations and discourse markers too: a marked relation is the one deviating from what can be inferred easily and by default. Recall the Bayesian formulation of the discourse relation inference process in 5.1. We decided that the effect of linguistic marking of a relation should be viewed as the likelihood term; thus:

$$markedness(r) \propto \frac{p(C,R)}{p(C) * p(R)} \tag{5.3}$$

In other words, the markedness of an instance of a discourse relation R, can be computed with respect to the information content of the cue C utilized in that instance. The logarithmic function of the above quantity is called the point-wise mutual information, which would be an ideal information theoretic measure to define the degree of the markedness of a relation made explicit by a discourse connective:

$$markedness(r) \propto \log \frac{p(C,R)}{p(C) * p(R)} = pmi(C,R)$$

Now, in order to compare relation senses with respect to how explicitly they are expressed by writers in a corpus of natural text, we can average over the markedness of all instances of each relation:

$$markedness(R) = \frac{1}{n} \sum_{i=1 \text{ to } n} markedness(r_i)$$
$$\propto \frac{1}{n} \sum_{i=1 \text{ to } n} pmi(c_i, R)$$

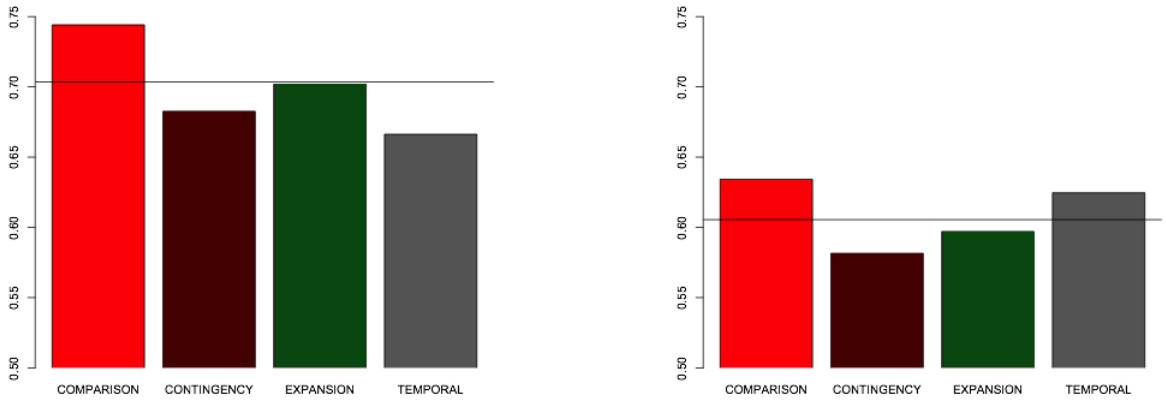
where n is the frequency of relations of the type R in the reference corpus, and c_i is the connective used in the ith instance of R. This formula gives us a value equal to the *connective use ratio* of R that we used in the previous section if we replace $pmi(c_i, R)$ with 1 in presence of the connectives and with 0 in absence of the connectives. In other words, the *connective use ratio* measure that we employed in the previous section's analyses is a binary version of markedness. The average markedness of a relation sense is indicative of the specificity of the connectives that are frequently used to signal it in the corpus. Using pointwise mutual information in the formula gives us a scalar but unbounded measure of markedness. It would be good to employ a normalized value between 0 and 1 to keep the numbers of this section in the same scale to that of the previous section. In order to do so, we use the normalization of pmi proposed by (Bouma, 2009) and scale it between 0 and 1 in the following way:

$$npmi(C, R) = \frac{pmi(C, R)}{-\log p(C, R)}$$
$$= \frac{\log \frac{p(C,R)}{p(C)p(R)}}{-\log p(C, R)}$$
$$= \frac{\log p(C)p(R)}{\log p(C, R)} - 1$$

$$0 < \frac{npmi(C, R) + 1}{2} < 1$$

$$\textit{avg. markedness}(R) = \frac{1}{n} \sum_{i=1 \text{ to } n} \frac{npmi(c_i, R) + 1}{2}$$

If a relation sense always co-occurs with a certain connective and that connective is not used in other relation senses, then the markedness reaches 1. On the other hand, relations



(a) Computed on explicit relations

(b) Computed on all relations

Figure 5.2: Markedness of the level-1 PDTB relations

that do not have specific connectives obtain a lower markedness score. Note that the general formula of markedness (as opposed to the binary version) is very sensitive to the sense labeling system that is employed in the annotation of the discourse relations. Two fine-grained relation senses that both co-occur often with a given discourse connective type would both gain a small markedness score. If another sense labeling scheme is used where the two relation senses are combined under a single coarser-grained class, this parent relation sense would gain a high markedness (if the connective does not co-occur together with any other relation in the new system).

For analyzing PDTB relations, we first calculate the markedness measure on the very coarse-grained relations. Considering only the explicit relations in the corpus which do have a connective gives us the numbers displayed in Figure 5.2a. The markedness of a relation instance by adapting this level of granularity is between 0.19 and 0.84 with an average of 0.70 over all explicit instances. If we want to include the implicit relations in our analysis, a null connective needs to be assumed for implicit instances which takes the same value across all relation senses. We ignore the inserted connectives by PDTB annotators in the implicit relations because the objective of the analysis is to determine the average markedness of the naturally generated relations. Figure 5.2b shows the markedness when calculated on all explicit and implicit relations in PDTB. This way we get values within the range of 0.20 to 0.80, and an average markedness of 0.61, which makes a quite difference with only considering the explicit population.

Figure 5.2a tells us that `COMPARISON` relations have their own set of cues, they pop out as the most marked relations meaning that they do not share their most frequent markers with other major classes of relations in the PDTB. Further experimental evidence also shows that these relations are more likely to cause processing difficulty than others when no connective is present, and that their markers have a more strongly disruptive

effect than other markers when used incorrectly in causal or additive relations (Murray, 1997). Under the information theoretic view, these observations can be interpreted as markers for COMPARISON relations causing a larger information update in their context. CONTINGENCY and TEMPORAL relations, on the other hand seem to have shared connectives since they obtain a lower markedness score. That proves to be true when these relations are investigated more closely. In particular, the frequent connectives *as*, *since* and *when* are distributed across TEMPORAL and CONTINGENCY relations or, in fact are labeled as marker of both relations in a single instance (see the ambiguity analysis in Section 3.2.3).

While CONTINGENCY relations turn out to have a smaller markedness score as we expected, the pattern illustrated by markedness measurement on explicit relations is not consistent with our predictions about the TEMPORAL relations (that they should be made explicit due to being discontinuous). In connective ratio analysis, we found TEMPORAL relations to be most often occurring with their explicit markers but now focusing on the information delivered by temporal connective types reveals that some of these cues are not as strong as we could call them discriminating features. Nevertheless, one important thing to note is that a lot of temporal connectives in the corpus are annotated with two relation senses (both relations have been marked at the same time). The way we count these relations (counting two instances each for one of the annotated relation senses), ends up giving us a lower markedness for TEMPORAL relations. Markedness of other relations are not affected by this artifact of the formula to the same extent, because in comparison to TEMPORAL relations, they are less often co-labeled with other relations.

Figure 5.2b gives us a complete picture of the markedness of the coarse-grained relations in PDTB. Since implicit and explicit relations are both considered, patterns become more similar to what we found in the connective use ratio analysis in the previous chapter. EXPANSION, which includes both continuous and discontinuous types, obtained a markedness close to average, TEMPORAL and COMPARISON including discontinuous and negative causal relations stand highest, and CONTINGENCY, which includes causal relations, remains the least marked class.

As we mentioned before, the markedness measure is very sensitive to the hierarchy of relation senses and how finer-grained types are put together or distinguished from one another. While a combined analysis over implicit and explicit relations would tell us about the overall informativity of the connectives used in relations of different types, we continue with analyzing explicit occurrences only to learn more about the finer-grained relation senses that share connectives vs. those being marked by specific connective types.

Figure 5.3 depicts the markedness calculated for explicit level-2 relations in PDTB. It differs significantly from what we see in Figure 5.2a. While in high level classification, COMPARISON relations are scored first in terms of markedness, the mid-level types of this relation sound to be sharing connectives. In fact, frequent markers of COMPARISON such as *but*, *however* and *although* are not highly specific regarding what fine-grained relation they mark. Contrarily, the level-2 relations under EXPANSION do have their own specific markers: *or, instead, unless* for Alternative, *except* for Exception and *for example, for instance* for Instantiation. As we explained before about the CONTINGENCY class, Condition relations have a very specific marker *if*, and Cause relations tend to share connectives with Asynchrony from the TEMPORAL class.

All in all, the markedness measurement reveals that some of the relations we merely categorized as explicit in our connective use ratio analysis are not as informative as others because connective types vary in terms of how specific they are to a given relation type. We now have a refined vision regarding the relation between the markedness of a relation in the corpus and its predictability given general cognitive biases. Causal relations are still more implicit than others. However, among continuous relations, some have highly informative connectives and some not. Negative polarity relations that we categorized as discontinuous are highly marked when a coarse-grained granularity is considered but are not distinguishable when it comes to finer categorization. This means that discontinuity as a general feature needs to become explicit but perhaps other dimensions that are different among sub-types of discontinuous relations require different levels of explicit marking.

Finally, an overall look at the values we obtain from the normalized pmi (and thus the markedness of the relations), makes it obvious that discourse connectives contain considerable amount of information about discourse relations. In the next set of experiments we will talk about other linguistic cues that contribute some information to the identification of relation senses, but to a much smaller extent compared to discourse connectives.

5.4 The effect of linguistic context

We so far found that the explicit but optional markers of discourse relations tend to be dropped to avoid redundancy in the more predictable relations with regard to prior expectations. This observation is in line with the general predictions of theories about efficient communication such as Grice's maxim of quantity (Grice, 1975) and the Uniform Information Density account (Levy and Jaeger, 2007): among meaning-equivalent formu-

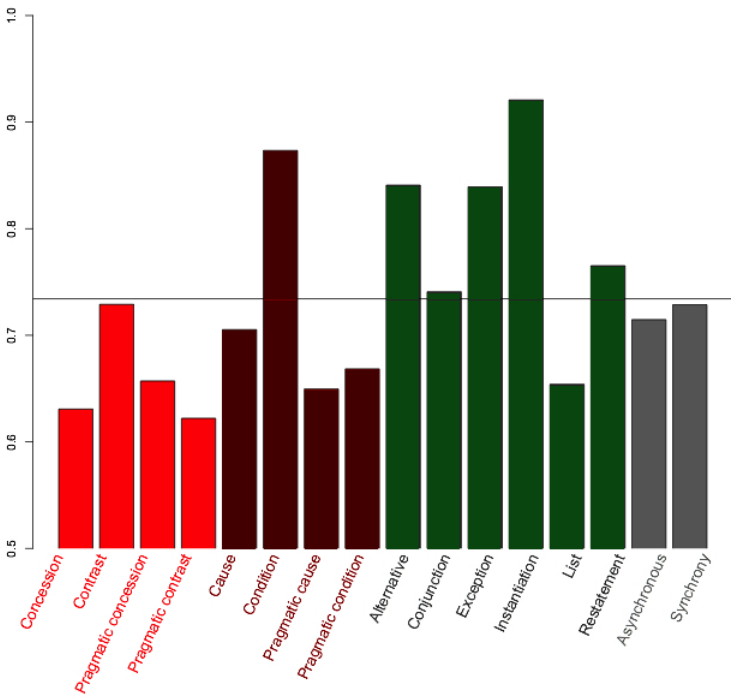


Figure 5.3: Markedness of the level-2 PDTB relations (explicit)

lations, speakers should prefer the one that encodes optimal amount of information, that is as much as is required for the reader to understand the message and not more than that. However, the more specific proposal of UID is that speakers should try to distribute information across an utterance uniformly so they make an optimal use of the communication channel capacity. This brings our attention to the effect of linguistic context in making a relation predictable. I propose that reduction of discourse connectives as optional markers of discourse relations should also correlate with presence of other linguistic cues in the context. The hypothesis is that discourse connectives should tend to be dropped more often when there are other strong cues in the context to predict the relation, and this should apply to production of both generally expected and unexpected relation types. Given that the language stimuli are perceived incrementally by the listener, connectives should be used to mark the relation when the first argument does not contain other relational cues. To test this hypothesis we need to detect the linguistic features of the sentences that are indicative of the relations they make with context.

Recently, a few systematic corpus studies have been conducted on the discovery of relational markers that are not traditionally considered as discourse connectives (Prasad et al., 2010; Das and Taboada, 2013; Duque, 2013; Webber, 2013). The obtained annotations throughout these studies provide evidence that, in fact, a lot of discourse relations benefit from other types of cues besides explicit discourse connectives. Examples include semantically related lexical chains, entity features, and morphological markers. Further-

more, the machine learning attempts for detection of implicit discourse relations that we reviewed in Chapter 2 reveal that lexical, syntactic and clause-level properties of the arguments to some extent help the identification of discourse relations (Pitler et al., 2009; Lin et al., 2009; Zhou et al., 2010; Park and Cardie, 2012; Rutherford and Xue, 2014). These observations motivate a computational investigation of the above hypothesis. In two experiments, we evaluate the effect of one potential predictor of `reason` relations in PDTB (Implicit Causality verbs) and one potential predictor of `chosen alternative` relations (negation markers), on presence/absence of the explicit connectives of these relations. The reason why we pick single linguistic features rather than employing an automatic discourse relation classifier for estimating the linguistic predictability of a relation is that the state-of-the-art accuracy of these classifiers at the fine-level classification is low. Also, we rather focus on a feature that is experimentally or empirically validated to be a predictor of a specific relation sense.

5.5 Experiment 2: connective reduction in presence of other cues

This section investigates the validity of the information theoretic explanation of connective reduction in presence of other incrementally available linguistic cues. Two Arg1 features of discourse relations have been selected to see whether or not, in their presence, connectives of certain relations are omitted. First, we look into the occurrences of Implicit Causality (IC) verbs, which have been identified in previous lab studies (Kehler et al., 2008; Rohde et al., 2006) as triggers for a `reason` continuation. The hypothesis is that explicit markers of `reason` relations, e.g., *because* and *since* should be more often omitted when Arg1 contains an IC verb than when other verbs are used. Second, we study negation words in Arg1 that have been identified by Webber et al. (1999) as downward entailing structures, which can license for the effect of connective *instead* in `chosen alternative` relations. The aim in this second experiment is to see how much the presence of negation in Arg1 affects the distribution of relations and, in turn, connective reduction in `chosen alternative` and possibly other relations that might as well be marked by negation.

5.5.1 Implicit causality verbs

IC verbs are a category of verbs that trigger specific expectation about the way different arguments of the verb or semantic roles in a sentence should be referred or

further explained in a causal continuation. IC verbs have been studied for many years in the context of reference (Caramazza et al., 1977; Koornneef and Van Berkum, 2006; Hartshorne et al., 2015). A set of studies have proven that these verbs trigger specific expectation regarding the type of discourse relation a sentence makes with its upcoming context (Kehler et al., 2008; Rohde et al., 2011; Rohde and Horton, 2014). In sentence completion tasks, Kehler et al. (2008) and Rohde et al. (2011) find that when the trigger sentence contains an IC verb people tend to provide a reason continuation for it, that is an answer to a *why* question (see a schematic example in (5-a)). Another class of verbs, called Transfer-Of-Possession (TOP verbs), on the other hand, tend to be continued with a theme of *what happened next* (like in (5-b).

(5) a. John *scolded* Mary. She had put thumbtacks on the teacher's chair.
b. John *shipped* Mary a package. She wrote him a thank you.

In order to see whether biases in production correlate with comprehension patterns, Rohde and Horton (2014) design a novel visual world experiment. Anticipatory looks to specific part of the screen are associated with expectation for either a reason or a transfer relation through a training task. People listen to 60 brief recorded passages that contain cause (equivalent of `reason` in PDTB) or consequence (equivalent of `result` in PDTB) coherence relations and a ball is rolled into one of the two pipes on the screen. After hearing each passage, participants are asked to guess which output pipe should return the ball (see Figure 5.4). Performance after training is then measured in a subsequent task where again for 24 items by hearing relations of either type participants should guess where the ball appears next. This is to make sure that the participants are able to associate relations in the stories with the visual pattern. Finally, 24 randomly mixed stories including verbs with different biases are played for listeners and their gazes are recorded to see which area on the screen is focused. Rohde and Horton find an early distinctive behavior right after hearing the cue in the utterances. In particular, subjects looked at the area associated with the reason relation as soon as they encountered the IC verb in the first sentence of the story. The patterns of anticipatory looks are very similar to when a strong discourse connective, i.e., *because* is used to mark the relation between two sentences. This indicates that people are able to take into account local cues like IC verbs to predict discourse relations early enough for the connective then to be a redundant marker.

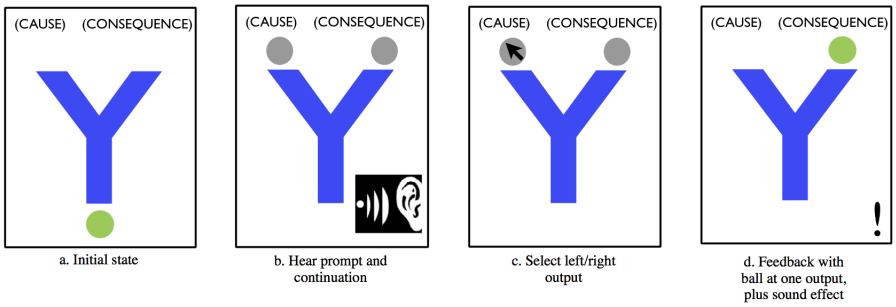


Figure 5.4: The training phase of the audio-visual experiment by Rohde and Horton (2014)

5.5.2 Predictions

The above findings have been obtained in very controlled laboratory settings. We do not know how far a relation can be anticipated when cues such as IC verbs occur in more complex naturally uttered sentences. A second question is concerned with the production of discourse connectives, that is whether speakers or writers are sensitive to the level of predictability introduced by IC verbs, and whether they would generate utterances that follow the efficient communication strategies. In order to address these questions, we conduct a corpus-based study to investigate patterns of production expected based on experimental findings and the UID account. In particular, by looking into the PDTB relations, we examine the following predictions:

- When a relation includes an IC verb in its Arg1, it should more likely be a `reason` relation than if it doesn't.
- When a `reason` relation includes an IC verb in its Arg1, then the connective should have a bigger chance to be reduces. In other words, among `reason` relations we expect a smaller ratio of connective use when an IC verb is present in Arg1 than in other cases.

Note that even if IC verbs are true markers of `reason` relations, they might not be as effective as they cause a connective reduction. Nevertheless, based on our UID-based argument, we expect a tendency towards connective reduction when other cues such as IC verbs exist in the context.

5.5.3 Data preparation

Implicit causality verbs need to be identified in PDTB relation to enable us test the above predictions. Dealing with natural and uncontrolled text in the corpus, requires us to have a comprehensive list of IC verbs so we make sure what is counted otherwise does not fall into the same category of verbs. We use a list of 300 IC verbs provided by Ferstl et al. (2011) which they annotated with fine-grained information regarding what argument of the verb is more salient. This data does not contain relational bias information which would be interesting for an scalar correlation analysis. Implicit and explicit relations in PDTB are examined to see if their Arg1 contained any of the verb types from Ferstl et al.'s list. To make sure that the IC verb worked as a cue in the sentence before encountering the connective, only relations with ordered arguments (Arg1-connective-Arg2) are considered. In total 1920 relation instances (about 5% of the data) are excluded in this filtration. Also, we only extract those instances where Arg1 is a single sentence and categorize the verb as either IC or non-IC. For this, the gold standard syntactic annotations from Penn Treebank are aligned and used together with relation annotations in PDTB. Table 5.3 contains statistics of the extracted data for our analysis.

	Total	IC verb in Arg1	
Implicit: reason relations	2462	153	(manually checked)
Explicit: reason relations	1324	96	(manually checked)
Implicit: all relations	15682	910	(automatically extracted)
Explicit: all relations	16147	1034	(automatically extracted)

Table 5.3: Total frequency of relations and the frequency of IC verbs appearing as the head of a single-sentence Arg1.

As a sanity check we look into the `reason` subset of the relations detected with an IC verb in their Arg1. From a total of 272 (164 implicit and 108 explicit), in 13 instances the verb is either a homonym of an IC verb (e.g., *to lie*) or some other unintended semantic sense (e.g., "leave it up to somebody" instead of "leave somebody"). Since such incorrectly tagged verbs are almost evenly distributed among implicit and explicit `reason` relations, they would not affect the connective use ratio, thus we do not worry about them in the analysis. Table 5.3 shows the manually checked numbers within `reason` relations. For other relations, only automatically extracted numbers are reported.

5.5.4 Analysis

First looking into the occurrences of IC verbs in relations of different types we find that a `reason` relation is significantly more likely if Arg1 contained an IC verb than when it did not ($p < 0.01$). However, the size of the effect is small: the likelihood of `reason` given an IC verb in Arg1 is 14.0%, and 11.9% given other verbs. This is support (though relatively weak) for IC verbs actually affecting the upcoming discourse relation in natural production of expository text.

The more interesting question in the context of the UID hypothesis, however, is whether markers for causal relations following IC verbs are more likely to be absent due to the higher predictability of the `reason` relationship. A comparison is made between the connective use ratio of `reason` relations where the Arg1 contains an IC verb to that of `reason` relations with non-IC verbs as the head of Arg1. As opposed to what we expected, the connective use ratio of `reason` relations with an IC verb in the Arg1 (39%) is not smaller but actually slightly larger than the connective use ratio of relations with other verbs in their first argument (35%). This result goes against our prediction, which can be due to any of the following reasons:

- There might be some hidden technical problems, e.g., a set of IC verbs that we counted as non-IC (because of the small size of our target list) might have not been uniformly distributed across relations of different types.
- It could be the case that IC verbs are not in principle as effective factors determining the relation in complicated sentences in expository text as they are in short narrative text.
- Finally, the result could mean that our UID-based account of connective omission does not apply. But this last conclusion would be too strong given the anyway small change in the likelihood of the relations by observing the IC verb.

Unfortunately, we do not have access to better IC verb lists (e.g., one of bigger coverage and including accurately measured biases of verb types for reason continuation). Also manual checking of all relations from the corpus would be an expensive task. Therefore, the first two mentioned possibilities cannot be easily ruled out, and without resolving them we can not argue much about the last point. Assuming that detection of IC verbs has been reliably done, the corpus data indicates that, in fact, IC verbs do not appear in production of `reason` relations as often as we can call them a strong cue of this relation.[4] Thus,

[4]The correlation between IC verb and relation being `reason` was only marginally significant, and this

they are not ideal linguistic features of Arg1 for us to examine our hypothesis regarding connective reduction in presence of contextual cues. In the next experiment a more reliably detectable linguistic feature is investigated.

5.5.5 Negation markers

Negative sentences are very frequent in natural text and their effect on linguistic inferences has been studied now for more than a decades (Lea and Mulligan, 2002; Staab, 2007; Schul, 2011; Orenes et al., 2014). In the context of discourse relations, sentence polarity has been traditionally used as one of the potential features for automatic identification of discourse relations. However, no concrete theory exists regarding which relations should be more likely to appear in continuation of a negative sentence. Webber's 2013 manual analysis of the `chosen alternative` relations is a recent focused study which points out a connective equivalent effect of negation in this particular type of relation. By looking into the explicit and implicit occurrences of *instead* in PDTB, as well as a self-collected set of *instead* sentences, Webber finds that the first argument of these relations often contains a downward entailing structure (like the verb *reject*) or explicit negation markers (*no*, *n't*, etc.). She also reports that the number of times negation markers appeared in implicit `chosen alternative` relations is bigger than the number of times they appear in explicit `chosen alternative` relations. This is an interesting observation which catches our eyes for the possibility of the UID-based connective reduction strategy functioning in `chosen alternative` relations. Fortunately, negation markers are relatively easy to detect automatically (when the scope does not matter), therefore it makes a suitable linguistic feature for a high coverage investigation of PDTB relations.

Another motivation for studying negation comes from the emphasized importance given to the relation polarity as a cognitively plausible dimension for classification of discourse relations (Sanders, 1997). While polarity of the relation is not necessarily equivalent to presence of explicit negation in arguments,[5] it is not unlikely that a correlation exists between the two. Now the point is that if negation turns out to be predictive of certain relations such as `chosen alternative` or in coarser granularity of negative

is the strongest correlation we found. If we calculate likelihoods of `reason` given IC with respect to the manually checked subset we get an even smaller likelihood, i.e., 12.8% that is closed to the likelihood of `reason` given other verbs 11.9%.

[5]Note, however, that negation in the surface is not necessarily equivalent to negative relational polarity. For example, "Mary loves John, but she pretends to ignore him." is a negative polarity relation without utilizing any covert negation, whereas "Mary doesn't love John and she pretends to ignore him" is a positive polarity relation including some negation.

polarity relations, then it would be an interesting Arg1 feature for a UID-based study of connective reduction.

5.5.6 Data preparation

Like in the previous experiment, all implicit and explicit relations which have a linear argument order (Arg1-connective-Arg2) constitute our target data. A binary feature is defined indicating whether any of the following negation words is present in an argument: {*not, n't, no, without, never, neither, none, non , nor, nobody, nothing*}. Table 5.4 is a summary of the statistics collected for implicit and explicit PDTB relations. Numbers in the table suggest that negative sentences are very frequent (but still considerably less frequent than positive polarity sentences). We furthermore see that the distribution of negation differs a bit between implicit and explicit relations, and each argument is different. Among all implicit and explicit relations under analysis about 14% turn out to have some negation in their Arg1.

Arg1 – Arg2	Explicit	Implicit	Total
POS – POS	14857	12155	27012
POS – NEG	1975	2153	4128
NEG – POS	2126	1758	3884
NEG – NEG	500	518	1018
	19458	16584	36042

Table 5.4: Distribution of negation in the arguments of PDTB explicit and implicit relations.

To test the reliability of the automatic procedure in discovery of negative words, we compared our list of explicit `chosen alternative` relations with the list manually analyzed by Webber (2013) and discovered only one difference where our algorithm found a negation in the first argument of the relation but it was not considered by Webber as a marker, or as she refers to that, it did not license the effect of the connective. Webber also detected five influential negation words in the attribution of the relations, as well as, five negations in a larger context rather than in the Arg1 boundaries. We do not consider such cases for the matter of consistency, i.e., we focus on the linguistic cues inside Arg1.

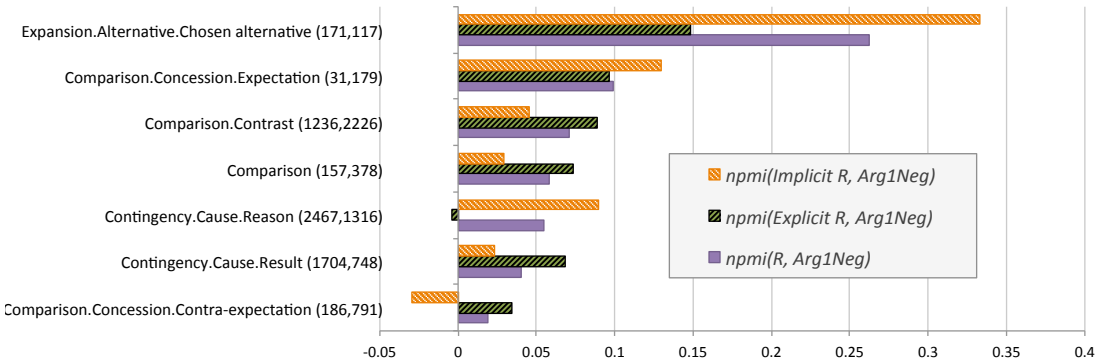


Figure 5.5: The npmi scores between relation senses and negation in Arg1

5.5.7 Predictions

In the analysis of the IC verbs, we looked into a specific type of relation, but negation markers are more frequent and might influence the distribution of relations more effectively. Besides the exploratory objective of this experiment to discover what relation types are marked by negation words, the following predictions are investigated:

- The presence of explicit negation words in a sentence increases the likelihood of a `chosen alternative` relation with the upcoming context.
- Among relation instances of a given type, like `chosen alternative`, that is considered to be marked by negation in Arg1, connective omission correlates with presence of the negation.

5.5.8 Analysis

In order to discover what relation senses are marked by negation in Arg1, we use the normalized pointwise mutual information as we did for measuring the effect of connectives on relations' markedness in Section 5.3.5. This enables us also to compare the contextual update by an explicit connective like *instead* and the other cue, i.e., negation in Arg1.

Relations marked by negation cues: In this analysis, all PDTB relations are considered with their fine-grained senses. Like in all experiments in this chapter, if a relation is annotated with more than one sense in the corpus, we count it for every sense separately. Figure 5.5 shows the relation senses obtaining a positive $npmi$ with the negation cue in Arg1. It reveals the set of relations in PDTB which are statistically marked by negation words in their Arg1. Frequency of implicit and explicit occurrences of every relation

sense are displayed in brackets.[6] Other relation senses either obtain a negative score, like `synchronous` which indicates that a negative polarity sentence in a text would least likely be followed by this relation, or a closed to zero score, i.e., no significant correlation. The `chosen alternative` relation, in particular, is located at the top, meaning that negation in Arg1 is highly predictive of this relation sense. Included in the graph are result of the calculation when implicit and explicit relations are considered separately. Considering only the implicit relations in the corpus reveals an even stronger pattern which is already a sign of interaction between the connective and the negation cues.

Figure 5.6 sorts relations based on the change in the likelihood of a relation sense, that is the proportion of the posterior probability after observing negation on the prior probability of the relation. The change of prior to posterior probability is significant based on a binomial test ($p < 0.001$) for the first six relations, and marginally significant for `contra-expectation` only when explicit instances are considered ($p < 0.05$). This means that after observing a negative sentence, the likelihood of the continuation is updated in favor of a `chosen alternative`, a subtype of `Concession`, a `Contrast` or a subtype of `Cause`.

The hunch we had regarding a correlation between sentence level negation and relation polarity applies to good extent. All frequent negative polarity relations in the corpus except the symmetric `Contrast` subtypes `opposition` and `juxtaposition` end up in our list of relations marked by Arg1 negation. The analysis confirms that the `chosen alternative` is the relation that benefits most from the negation cue, thus would be a good candidate for investigation of the connective reduction hypothesis.

Connective reduction: Remember in Section 5.3.5 the average markedness of the relations when calculated based on the $npmi$ of the discourse connectives was always higher than 0.5. The $nmpi$ scores obtained for the negation features are much smaller compared to the average $npmi$ of discourse connectives. In particular, for `chosen alternative` which obtains the highest $npmi$ with negation cues (0.4), we have the connective *instead* with an $npmi$ of 0.8. Nevertheless, we still expect if UID applies to discourse connective omission, the presence of negation cues in a given instance of this relation should bias the speaker towards dropping the connective. I conducted separate correlation analyses between presence/absence of negation and presence/absence of the connective for all relations that we found were marked by negation. The results are as follow:

- Among the `chosen alternative` relations, absence of the discourse connec-

[6] Only labels that have 30 or more implicit and 30 or more explicit instances in the corpus are displayed.

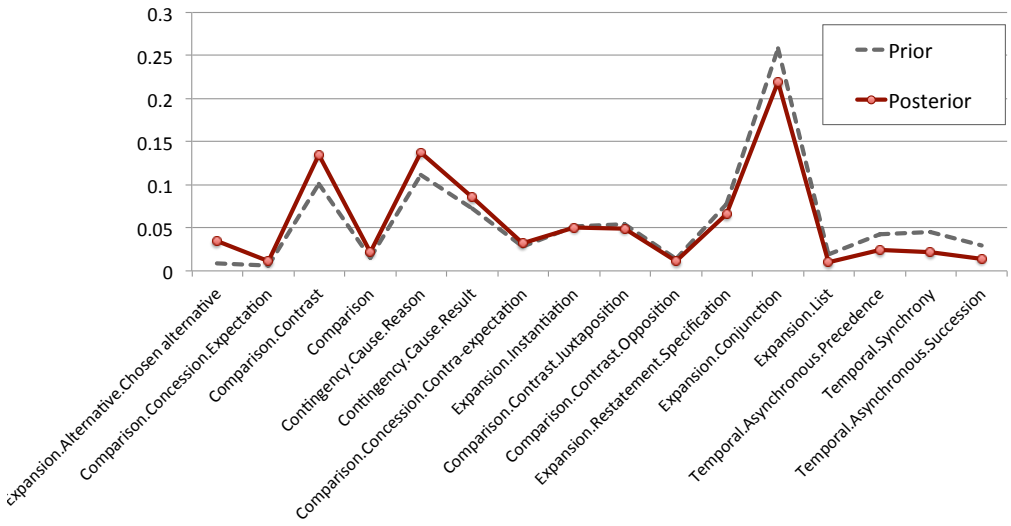


Figure 5.6: Change of likelihood after observing the negation cues

tive is positively correlated with the presence of negation: the connective use ratio is only 24.4% for relations with some negation in Arg1, whereas it is 61% for the rest of `chosen alternative` relations ($p < 0.001$).

- Figure 5.5 indicates that two other relation senses, i.e., `expectation` and `reason` show similar patterns. They also tend to be marked more strongly by negation when only their implicit occurrences are considered. The effect is significant for `reason` and not for `expectation` (note that only 31 implicit instances of this relation exist in the corpus).

- `COMPARISON` and `COMPARISON.Contrast` relations show an opposite trend, i.e., while negation in Arg1 increases the likelihood of these relations, the connectives marking these relations tend to be dropped in the presence rather than absence of the negation feature. This effect is not significant after all.

These observations confirm our UID-based prediction about the `chosen alternative` relation, in particular: writers tend to omit connectives in this relation when a strong contextual cue, namely, a negation marker exists in the Arg1. Generalizing this to other relations with lower degree of markedness by negation words does not seem to be easy. One reason is the less salient a linguistic feature in a relation, the less we could argue how it affects the use of explicit connectives. This makes sense computationally too. We proposed that the discourse level UID should be viewed in terms markedness changing in a small range across occurrences of a relation type in a corpus. Thus, linguistic features or cue elements that do not contribute much to the markedness of a relation should have less correlations. In case of negation and connective *instead*, both are fairly good markers of the `chosen`

`alternative` relation thus we expect a clear correlation. Similar pairs of cues (in terms of informativity) should be studied when other relations are under investigation.

5.6 Summary

The observations that we made throughout this chapter add to our knowledge of when a discourse connective may or may not be produced in natural text generation. We found a major effect of a set of **prior expectations for specific relational continuations** on deciding what relation senses in general need more explicit marking. Causal and continuous relations that are expected by readers in exposure to consecutive sentences tend to be left implicit in natural text. This can be interpreted as a sign of communication principles being observed and considered by writers subconsciously: we have to make a relation explicit if our reader does not expect it by default.

A minor effect of the **linguistic context** also came out in our study of the chosen alternative relations. These relations can be expressed by a very informative connective *instead*. In a considerable portion of chosen alternative relations, we also see some type of negation in the first argument. We showed that negation in Arg1 is a statistically licensed cue for detecting chosen alternative relations, and more interestingly if this implicit feature is present, then the connective is more likely to be left out. The same pattern applies to other relations whose likelihood is increased significantly by the negation cue being present in their Arg1, but the effect size depends on both the informativity of the negation cue and the informativity of the connective being omitted.

These observations provide initial support for an information theoretic approach to discourse relation marking. In summary, we showed that the predictability of a relation given the default expectations at the reader side and/or given the local linguistic features is an important factor in discourse-level production. To the best of our knowledge, this is the first psycholinguistic explanation of the way speakers connect pieces of a discourse for it to be understood by their audience without being over-informative. Taken together our findings support an account of language processing that views comprehension and production as mirror images, meaning that the two processes constrain each other. We found that the predictabilitiy of a relation from the view point of a listener determines the way the relation is encoded by the speaker. If production of discourse relations was independent from their comprehension, we should not have observed correlation between predictability of a relation and its form. This suggests that the speakers have a model

for their discourse-level production with some flexibility that relates to the listener-side constraints.

All experiments in this chapter have been conducted on large scale text from WSJ. As we saw, for example in our study of Implicit Causality verbs, dealing with uncontrolled heterogeneous sentences makes it more difficult for a corpus-based study to come up with nice and clean measurements and concrete conclusions. Experimental studies need to be conducted on comprehension and production of discourse relations to confirm the hypotheses we proposed and investigated in this chapter. At the same time, finding patterns pertaining to very fine levels of informativity at the discourse/pragmatics level in this type of data is already an encouraging achievement. This is a proof that psycholinguistic theories can be tested on natural data, not only on carefully designed stimuli in a laboratory experiment.

Chapter 6

Conclusion

This chapter summarizes the thesis and presents several directions for future work. The major findings and contributions of this work to both theory and methodology are highlighted.

6.1 Summary

The goal of the thesis was to establish a new framework for studying discourse relations and their markers from the viewpoint of communication and information theory. In Chapter 2, we reviewed previous work on discourse coherence and elaborated the importance of the discourse relations as building blocks of a coherent text and analytic concepts for explaining human communication when inferences are involved. The framework we proposed in Chapter 3 is based on previous theories of how information is transferred from a speaker to a listener. Previous works in this domain have looked into other levels of sentence processing, i.e., phonology, morphology, syntax and semantics within the boundary of sentences. In this framework, every unit or symbol in an utterance has some information content that can be measured.

Firstly, we proposed that the information content of a discourse marker should be formulated with respect to the ambiguity it removes in relational interpretation. In order to show how, we used the annotation of discourse relations and their markers, particularly, discourse connectives in Penn Discourse Treebank and employed a set of established measures in information theory to model the meaning of connectives. The general analysis of the corpus in Chapter 3 revealed significant differences among discourse connectives in terms of the type and amount of relational information they deliver. In particular, we identified three types of ambiguities: a connective can mark two relations at a time (*while* can mark temporal synchronous and contrast relations between events), or it can mark different relations depending on the context (*since* sometimes marks a temporal relations and sometimes a causal relation), or it can mark multiple sub-types of a general class of relations (*but* can be used in contrast or concessive relations which are sub-types

of comparison or negative polarity relations). Connectives of all these types can be identified by our quantified methodology, as relations with multiple senses exhibit smaller information content compared to highly specific discourse markers (e.g., *instead* as a marker of chosen alternative relations). We focused on the third type of ambiguity for a more detailed examination.

We designed and ran a set of experiments on *but* and *although* as a case study of multi-sense connective to show that the information content of the connectives calculated within the proposed framework can be discussed and compared in a quantified manner. This set of experiments, presented in Chapter 4, confirmed that the usage patterns of a connective collected from the reference corpus of natural text, i.e., the distribution of the relations that co-occur with the connective can predict how the connective is interpreted in a new context. More specifically, we found that in contrary to an underspecified account of multi-sense connectives (Fraser, 1999), there is very fine-grained information encoded in these linguistic elements that bias the listener towards specific interpretations. While *but* and *although* are both from the same major class of discourse connectives, they generate different expectations in the same context. Even *although* in alternative arrangements (arg1-although-arg2 vs. Although-arg2, arg1.) generates different interpretations and expectations regarding the way the story should be continued. All effects that we observe throughout our experiments are predictable by looking into the distribution of the discourse relations these connectives mark in the corpus.

Finally, in our second set of experiments in Chapter 5, we focused on the production of discourse connectives by raising the question of implicitness: when are discourse connectives utilized for making relations explicit and when are they reduced? The Uniform Information Density theory (Levy and Jaeger, 2007) was introduced, which proposes that the optional markers in a language should tend to be dropped in contexts that they can be easily predicted. We applied this idea to discourse connectives in English which are syntactically optional in most contexts. We predicted that discourse connectives as markers of semantic discourse relations, should tend to be reduced when a relation is predictable given either a listener's prior expectations or other linguistic cues in the context. This hypothesis was tested through our second set of experiments. We conducted a large-scale corpus-based study of implicitly and explicitly marked discourse relations in PDTB to examine the correlation between connective reduction and relation predictability. Prior expectations and other linguistic cues of discourse relations were identified based on previous cognitive science theories, psycholinguistic experiments and linguistic studies of discourse relations. According to our analysis, causal and continuous relations, which

based on established theories are expected by default, tend to be expressed without discourse connectives. On the other hand, unexpected types such as concessive and temporally discontinuous relations show a higher degree of linguistic markedness or use of discourse connectives. Regarding the effect of other linguistic cues, we looked into specific markers in Arg1 of the relations that were previously examined for their online effect on inferences: implicit causality verbs and negation markers. We collected relations of different types in the corpus that included these linguistic cues and hypothesized that in the predictable relations we should see larger proportion of connective reduction. Implicit causality verbs did not turn out to be a strong predictor of the `reason` relations in the corpus contrary to what we expected based on previous experimental studies (Rohde and Horton, 2014); and they did not correlate with presence/absence of the optional markers for `reason` relations (e.g., *because*). However, negation turned out to be a strong marker of a set of relations and in particular the `chosen alternative` relations, in line with previous work (Webber et al., 1999). In this type of relation, presence of a negation marker in the first argument correlates significantly with presence/absence of the optional discourse connective *instead*. This suggests that an increase in relation predictability corresponds with a decrease in the likelihood of explicit marking.

Our findings provide some positive evidence for an account of connective omission in predictive context. From a more general perspective, all of our experiments put together support the hypothesis of an information theoretic approach to communication: speakers try to encode information in a way that is easy to process and enough for the right interpretation while being efficient. Listeners also decode massages by maximal use of the information made available to them, and they experience processing difficulty when the input does not follow their expectations, which are shaped based on previous exposure to natural linguistic stimuli.

6.2 Contributions

The thesis includes a comprehensive review on discourse relations and their linguistic markers. Chapter 2 looked into the more linguistic and technical studies, involved with definition of discourse relations, annotated corpora and the linguistic features that are used in discourse parsing and automatic identification of relation senses in text. Throughout the introductory part of Chapter 4, we reviewed the psycholinguistic studies on comprehension of discourse relations that were focused on specific discourse connectives. In addition to the survey of previous work across disciplines, the following contributions have been

made to the fields of computational linguistics, psycholinguistics and natural language processing.

6.2.1 Semantic representation for discourse markers proposed

From a formal semantics perspectives, discourse connectives do not refer to concepts and most of them do not affect the truth-conditional state of the utterances they connect. This makes it difficult to define what a discourse connective like *but* means, or how it differs from other connectives like *because* or *although*. The typical practice in previous theoretical work has been to focus on a discourse connective and define its function by exemplifying the different contexts where it can be used. This is a purely descriptive approach. We proposed that discourse connectives can be viewed from an information theoretic perspective and the information they contain can be approximated based on their distribution across discourse relations of different types in natural text. This way we take the descriptions we need for explaining the meaning(s) of a connective from discourse relations, and calculate a distributional vector for the connective, representative of its semantics.

The probabilistic approach has the following advantages over the descriptive one. Different types of ambiguous vs. specific discourse markers can be identified by calculation of the information content. In this framework, we can easily explain how an ambiguous or multi-sense connective biases a reader towards one among several possible interpretations (relations) that are inferable from the context. We can also compare two discourse connectives based on their distributional vectors. Two connectives that have similar distributions across relations of different types in a reference corpus are similar in meaning, whereas connectives with different distributions are different in meaning (see an abstract comparison in Figure 6.1).

6.2.2 Multi-sense connectives investigated

In Chapter 4, we focused on the inferences that are the mutual product of a discourse connective and the content of both relational arguments in a short narrative text. We coupled our corpus-based analysis of the meaning of *but* and *although* with an offline coherence judgement test and an online eye-tracking reading experiment to see how different possible interpretations of the connected sentences were triggered by the connective. A comprehensive study of this type on discourse relations has not been carried out before.

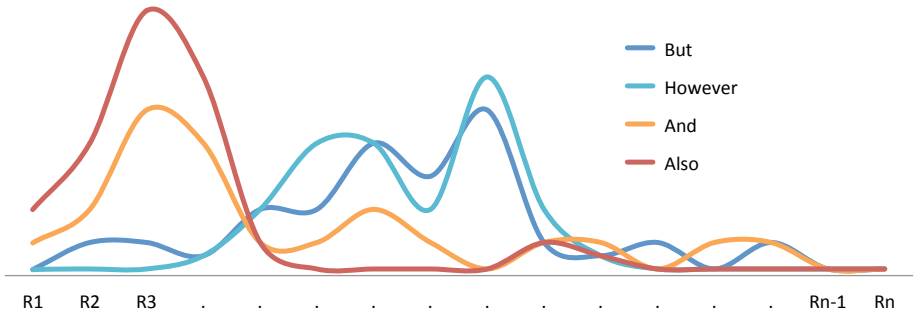


Figure 6.1: Abstract view of the connectives with similar or different relation distributions

Based on the corpus study, we chose very closely related connective types to capture fine-grained differences in the resulting inferential processes. Our experiments suggest that a distributional account of connectives' function is more accurate than classic approaches in predicting the way multi-sense connectives guide relational inferences. Previous experimental studies on discourse connectives have been primarily focused on the local effect of these markers, i.e., on integration and prediction of an immediately attached discourse segment. Also, when pairs of connectives were compared, they were usually chosen from very different categories (e.g., the causal *therefore* vs. the concessive *however*). Such a setup usually makes one connective incoherent in the context where the other connective fits. In contrast, our experiments aimed to show that two relatively similar connectives like *but* and *although* might be used coherently in the same context, but each generates an inference that is significantly different from the other. We found that different inferences cause different expectations regarding the way a story should be continued content-wise.

6.2.3 Question of connective reduction raised

This thesis represents the first work examining why some relations are marked explicitly while others are left implicit. The topic has so far attracted the attention of researchers in psycholinguistics and natural language processing. We explained how previous theories on efficient communication strategies could provide a hypothetical answer to this question: connectives should be dropped in predictive context to avoid redundancy and keep the information density uniform. We identified and discussed the effective factors on predictability of discourse relations and conducted a large-scale study on English connective use/reduction by looking into implicit and explicit relations in PDTB. According to what we found, connectives tend to be reduced not only in generally expected relation types but also when the relation they mark is predictible in a specific linguistic

context. We also defined a measure of markedness for a relation regarding the information contributed by a utilized connective to make it explicit. Using the ideas and methodology developed as part of this thesis, similar studies have been conducted on other languages and resources. In particular, Jin and de Marneffe (2015) looked into a corpus of Chinese and found similar patterns of relation markedness as that of the English ecquivalents. Hoek and Zufferey (2015) and Hoek et al. (2015) conducted a set of cross-lingual analysis on implicitation, which revealed that translation of discourse relations sometimes occurs with reduction of the discourse connectives and it happens more often when the relation is an expected type. Our work also motivated research on automatic identification of the type of context, where a discourse connective should be used vs. dropped. For example, Patterson and Kehler (2013) trained a binary classifier on implicit and explicit relations in PDTB and found that a set of automatically extracted linguistic features from the arguments of a discourse relation and its larger context can determine whether a connective should be used or not. Yet, their method is not explanatory regarding why each relation instance is classified as either case. Our experiments on the linguistic context, in particular, the study of negation and discourse connectives, revealed that the two types of markers tend to occur exclusively in `chosen alternative` relations. Thus, the explanation we proposed for why a relation should be expressed by a connective in one context and not in another is that speakers try to avoid redundancy at the level of discourse relations. This finding also has an important implication to the NLP attempt for identification of implicit relations: implicit and explicit instances of a relation might be inherently different regarding the content and surface features in the arguments, because discourse connectives are naturally used when the relation is not predictable given the context or not expected by the reader. An attempt to overcome data sparsity in implicit relation identification has been to use explicit relations from unlabeled data. This approach assumes that implicit and explicit relations of a given type are similar in terms of other features, thus tries to leverage connectives in unannotated text to harvest more training data for identification of implicit relations. Previous experiments provide positive and negative evidence for functionality of this method (Marcu and Echihabi, 2002; Pitler et al., 2009; Sporleder, 2008; Zhou et al., 2010; Hernault et al., 2011; McKeown and Biran, 2013). If our theory about the efficient use of discourse connective applies to natural text, learning from explicit relations to identify implicit relations should not be easy if ever possible. We believe that the similarity between implicit and explicit relations must locate in very abstract semantic levels rather than at the level of easily detectable surface features. This is because the latter is in more control of the speaker and avoiding redundancy should naturally result in the reduction of surface cues.

6.2.4 Theories of communication examined

The information theoretic approach to human communication has a long history, but the present thesis is the first comprehensive work in this domain that looks into discourse relations. We developed a framework to explain comprehension and production of discourse relations in a quantified manner and relate them to other levels of language processing. Regarding comprehension, we took the first steps towards formulation of a discourse marker's surprisal, a measure that is used in computational psycholinguistics to relate processing difficulty of a linguistic unit with respect to its predictability in discourse. Currently available annotated corpora are too small for implementing a generative discourse parser and building a computational model of processing load. Nevertheless, our study of *but* and *although* indicates that the information content of a discourse connective directly affects comprehension of multi-sentence text. Regarding production, we proposed that the Uniform Information Density theory that successfully explains a variety of patterns in spoken word duration and articulation (Buz et al., 2014; Demberg et al., 2012; Sayeed et al., 2015), morphology (Kurumada and Jaeger, 2013), syntax (Jaeger, 2010), lexical choices (Piantadosi et al., 2011; Mahowald et al., 2013) and referring expressions (Tily and Piantadosi, 2009; Kravtchenko, 2014), can also explain some patterns in discourse marker production. As UID and other theories on efficient communication predict, natural text reveals a tendency in speakers to choose the shorter form (drop the explicit cue) when a relation is expected or would be easily processed by the listener without the presence of the connective. Our findings regarding the connective reduction patterns in natural text also provide empirical evidence for the continuity and causality-by-default hypotheses (Segal et al., 1991; Murray, 1997; Sanders, 2005) telling that the events being narrated in consecutive sentences are expected to have causal and continuous relations unless otherwise is explicitly marked.

6.3 Future work

While we examined the information theoretic account of discourse relations through a series of experiments, our work raises new questions and presents several promising avenues for future research in both application-oriented and theoretical domains. This section presents the major future work directions.

6.3.1 Application-oriented research

Learning the meaning of discourse connectives from text would be one of the future directions for application-oriented research. Our study showed how important discourse relations are in defining the meaning of discourse connectives. Most connectives are not bound to a single sense and can be used in a variety of contexts, but context for a discourse connective should be viewed in terms of the discourse relation(s) in which the connective is utilized. Work on distributional semantics has not yet explored discourse cues in much depth. Hutchinson (2005) conducted a set of experiments regarding the possibility of representing the meaning of a discourse connective based directly on its surrounding words. However, representation of the context in this work does not pertain well to discourse level semantics. For example, instead of word-pairs that co-occur with a connective they collect unigrams from each argument of the connective. A set of recent work on identification of implicit discourse relations (Rutherford and Xue, 2014; McKeown and Biran, 2013; Braud and Denis, 2015) use more relevant distributional vector representations for relation identification that can be adapted to the computational modeling of the meaning of discourse connectives. Such a model can then be used as a component in any system that aims to generate multi-sentence text (e.g., summarization and dialog systems), extract meaning from multi-sentence text (e.g., search engines and question answering systems), or evaluating the coherence of multi-sentence text (e.g., readability scoring and learner assessment).

The second place for research on the generation of natural sounding discourse is the detection of the contexts in which a relation needs to be made explicit. The straightforward approach based on our theory would be to see if enough information about the relation is already encoded in the two arguments, i.e., two sentences that are intended to convey a specific relation. While redundancy exists in natural languages, discourse connectives are not used arbitrarily. These elements tend to be used when the relations they mark are less predictable in the context or less expected by the listeners. Some related studies on adjusting informativity and redundancy avoidance can be found in the literature on coreference chains (Tily and Piantadosi, 2009; Kravtchenko, 2014). Similar to the way that using the right form of a referring expression in a mutli-sentence text makes it more natural, the right way of marking discourse relations contributes to the coherence of the text. Not only selecting the right marker for places where a relation cannot be inferred from the context is necessary, but also reducing the redundant markers where they would not add any new information can result in a more coherent discourse.

The following examples compare the natural use of referring expressions and discourse connectives in a story taken from an English tutoring website with a superfluous use of these markers in an artificial (modified) version of the story.[1]

(1) Referring expressions

a. In a huge pond, there lived many fish. They were arrogant and never listened to anyone. In this pond, there also lived a kind-hearted crocodile. He advised the fish, "It does not pay to be arrogant and overconfident. It could be your downfall." But the fish never listened to him. (original)

b. In a huge pond, there lived many fish. **The fish** were arrogant and never listened to anyone. In this pond, there also lived a kind-hearted crocodile. **The crocodile** advised the fish, "It does not pay to be arrogant and overconfident. **Being arrogant and overconfident** could be your downfall." But the fish never listened to the **crocodile**. (modified)

(2) Discourse markers

a. The crocodile heard all this. When the fishermen left, he slowly slipped into the pond and went straight to the fish. "You all had better leave this pond before dawn. Early morning those two fishermen are going to come to this pond with their net, " warned the crocodile. (original)

b. The crocodile heard all this, **so** when the fishermen left, he slowly slipped into the pond and went straight to the fish, **then** he warned "You all had better leave this pond before dawn **because** early morning those two fishermen are going to come to this pond with their net". (modified)

Experimental studies on particular tasks (e.g. summarization) could explore whether this information theoretic account of discourse coherence can enhance the quality of generated texts in practice.

[1] Stories taken from www.english-for-students.com.

6.3.2 Theoretical research

This was the first corpus-based study of communication principles and information theory at the level of discourse relations. Our findings altogether indicate an interaction between production and comprehension mechanisms that is reflected in language data. The extent of this interaction remains unknown. Recent experimental studies on related phenomena, such as the production and comprehension of referring expressions provide evidence that production can sometimes be insensitive to semantic biases that affect comprehension (Rohde and Kehler, 2014). In one of our experiments, we investigated the presence of implicit causality verbs in Arg1 of the causal relations and found no interraction between that and the reduction of the discourse connectives. Working with large-scale natural text makes it difficult to control for possible confound factors, such as the length of the relational arguments, the place where connectives appear, and contextual linguistic features that are not considered in the analysis. Conducting a set of controlled sentence production experiments would be one way to extent our study for more clear results.

The corpus-based study of *but* and *although* revealed that, with regard to the PDTB annotation schema, both connectives are used in a variety of relations. However, each connective has a dominant bias in terms of the frequency of its coocurrence with relations of specific types. Our comprehension experiments showed that distributional differences result in different interpretations when alternative connectives are used in identical contexts. Inferential effects might not be obvious if we only look at the text span including the connective but they become important when the larger context is encountered. While in the offline coherence judgment experiments we captured the differential effects of *but* and *although*, the results of our online reading experiment were rather weak. This might be due to the sensitivity of eye-tracking measures to the type of stimuli. Future experiments with more carefully designed stimuli and perhaps incorporating other methodologies (e.g., EEG, self-paced reading and visual world paradigm) might help us better understand the online processing of a new sentence after inferring a particular discourse relation triggered by a connective.

One of our initial objectives was to provide a framework for calculation of surprisal, a widely used measure for modeling processing difficulty in human sentence comprehension. Surprisal is calculated based on the probability of a word given its context. We elaborated how relational context can be defined for a word. However, given the type and size of available data annotated with discourse relations and markers, implementation

of a surprisal model covering discourse level dependencies would be very challenging. Powerful implementations of surprisal use highly accurate syntactic parsers and semantic role labeling systems. The accuracy of state-of-the-art discourse parsers are too low for an appropriate modeling of discourse-level processing — 42.72 overall accuracy in relation sense identification, according to the CoNLL 2015 shared task (Xue et al., 2015).[2] Thus one direction for the future work would be to enhance these parsers. In particular, high-level semantic features need to be defined and extracted from text for likelihood calculation of relations between neighboring sentences before we can proceed with modeling discourse-level surprisal.

Since our studies have been all focused on English and specific corpora, it also leaves room for similar investigation on different languages and other types of text.

6.4 Closing remarks

Modeling human language processing is an attractive topic of research for scientists in different fields such as linguistics, artificial intelligence and cognitive science. The fundamental questions about human understanding of events and relations between them go back to ancient philosophy. What shapes human expectations when they hear a story, what surprises them and what is easy for them to infer despite a lack of explicit mention, are all the high-level questions that motivated this thesis. It is a modest contribution to our understanding of particular phenomena in language processing and leaves a lot of open questions to be answered in the future. Development of larger corpora of discourse relations with comprehensively annotated linguistic features as well as machine learning methods for harvesting unlabeled data for automatic detection of discourse relations would provide a better framework for examining cognitive theories about discourse processing. Understanding the comprehension and production mechanisms in human communication would also provide knowledge for design and implementation of task-oriented applications. Thus findings on the relevant topics should be communicated between the two communities of researcher. This principle was considered in the current study. We hope to see more interdisciplinary work in the future.

[2] http://www.cs.brandeis.edu/ clp/conll15st

List of Figures

List of Tables

Bibliography

Asher, N. and Lascarides, A. (1998). Bridging. *Journal of Semantics*, 15(1):83–113.

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Asr, F. T., Sonntag, J., Grishina, Y., and Stede, M. (2014). Conceptual and practical steps in event coreference analysis of large-scale data. In *Proceedings of ACL, the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Baltimore, Maryland, USA*.

Bach, K. (1999). The myth of conventional implicature. *Linguistics and philosophy*, 22(4):327–366.

Bach, K. (2006). The top 10 misconceptions about implicature. *Drawing the Boundaries of Meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, pages 21–30.

Blair-Goldensohn, S., McKeown, K., and Rambow, O. (2007). Building and refining rhetorical-semantic relation models. In *HLT-NAACL*, pages 428–435.

Blakemore, D. (1987). Semantic constraints on relevance.

Blakemore, D. (1992). *Understanding utterances: An introduction to pragmatics*. Blackwell Oxford.

Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*, volume 99. Cambridge University Press.

Blühdorn, H. (2008). Subordination and coordination in syntax, semantics and discourse. *Subordinationversus Coordinationin Sentence and Text. A Cross-Linguistic Perspective, Amsterdam*, pages 59–85.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, pages 31–40.

Braud, C. and Denis, P. (2015). Comparing word representations for implicit discourse relation classification. In *Empirical Methods in Natural Language Processing (EMNLP 2015)*.

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

Bruder, G. A., Duchan, J. F., and Rapaport, W. J. (1986). *Deictic centers in narrative: An interdisciplinary cognitive-science project*. State University of New York (Buffalo). Department of Computer Science.

Bunt, H., Prasad, R., and Joshi, A. (2012). First steps towards an iso standard for annotating discourse relations. In *Proceedings of the Joint ISA-7, SRSL-3, and I2MRT Workshop on Semantic Annotation and the Integration and Interoperability of Multimodal Resources and Tools*, pages 60–69.

Buz, E., Jaeger, F., and Tanenhaus, M. K. (2014). Contextual confusability leads to targeted hyperarticulation. In *Proceedings of the 36th annual conference of the cognitive science society*.

Canestrelli, A. R., Mak, W. M., and Sanders, T. J. (2013). Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Language and Cognitive processes*, 28(9):1394–1413.

Caramazza, A., Grober, E., Garvey, C., and Yates, J. (1977). Comprehension of anaphoric pronouns. *Journal of verbal learning and verbal behavior*, 16(5):601–609.

Carlson, L., Marcu, D., and Okurowski, M. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Current and new directions in discourse and dialogue*, pages 85–112.

Caron, J., Micko, H. C., and Thuring, M. (1988). Conjunctions and the recall of composite sentences. *Journal of Memory and Language*, 27(3):309–323.

Carrell, P. (1982). Cohesion is not coherence. *TESOL quarterly*, pages 479–488.

Cozijn, R., Noordman, L. G., and Vonk, W. (2011). Propositional integration and world-knowledge inference: Processes in understanding because sentences. *Discourse Processes*, 48(7):475–500.

Das, D. and Taboada, M. (2013). Explicit and implicit coherence relations: A corpus study.

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, 8(8):1117–1121.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Demberg, V., Sayeed, A. B., Gorinski, P. J., and Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367.

Demberg-Winterfors, V. (2010). Broad-coverage model of prediction in human sentence processing.

Drenhaus, H., Demberg, V., Köhne, J., and Delogu, F. (2014). Incremental and predictive discourse processing based on causal and concessive discourse markers: Erp studies on german and english.

Duque, E. (2013). Signaling causal coherence relations. *Discourse Studies*, page 1461445613496358.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4):491–505.

Fellbaum, C. (1999). *WordNet*. Wiley Online Library.

Feng, V. W. and Hirst, G. (2012). Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 60–68. Association for Computational Linguistics.

Ferstl, E., Garnham, A., and Manouilidou, C. (2011). Implicit causality bias in english: a corpus of 300 verbs. *Behavior Research Methods*, 43(1):124–135.

Fillmore, C. J., Johnson, C. R., and Petruck, M. R. (2003). Background to framenet. *International journal of lexicography*, 16(3):235–250.

Fraser, B. (1990). An approach to discourse markers. *Journal of pragmatics*, 14(3):383–398.

Fraser, B. (1998). Contrastive discourse markers in english. *PRAGMATICS AND BEYOND NEW SERIES*, pages 301–326.

Fraser, B. (1999). What are discourse markers? *Journal of pragmatics*, 31(7):931–952.

Frazier, L. and Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2):178–210.

Grice, H. P. (1975). Logic and conversation. *Reprinted in Studies in the Way of Words 1985*, page 2240.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Haberlandt, K. (1982). Reader expectations in text comprehension. *Advances in Psychology*, 9:239–249.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8.

Hall, A. (2004). The meaning of but: A procedural reanalysis. *UCL Working Papers in Linguistics*, 16:199–236.

Hall, A. (2007). Do discourse connectives encode concepts or procedures? *Lingua*, 117(1):149–174.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman (London).

Hartshorne, J. K., Sudo, Y., and Uruwashi, M. (2015). Are implicit causality pronoun resolution biases consistent across languages and cultures? *Experimental psychology*.

Hernault, H., Bollegala, D., and Ishizuka, M. (2011). Semi-supervised discourse relation classification with structural learning. *Computational Linguistics and Intelligent Text Processing*, pages 340–352.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive science*, 3(1):67–90.

Hoek, J., Evers-Vermeul, J., and Sanders, T. J. (2015). The role of expectedness in the implicitation and explicitation of discourse relations. *DISCOURSE IN MACHINE TRANSLATION*, page 41.

Hoek, J. and Zufferey, S. (2015). Factors influencing the implicitation of discourse relations across languages. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11)*, page 39.

Hovy, E., Mitamura, T., Verdejo, F., Araki, J., and Philpot, A. (2013). Events are not simple: Identity, non-identity, and quasi-identity. *NAACL HLT 2013*, page 21.

Hovy, E. H. and Maier, E. (1992). Parsimonious or profligate: how many and which discourse structure relations? Technical report, DTIC Document.

Hume, D. (1784). *An Enquiry Concerning Human Understanding*. New York: The Liberal Arts Press, 1955 edition.

Hume, E. (2004). Markedness: A predictability-based approach. In *Annual Meeting of the Berkeley Linguistics Society*, volume 30.

Hutchinson, B. (2005). The automatic acquisition of knowledge about discourse connectives.

Iten, C. (2000). Although revisited. *Working Papers in Linguistics*, 12.

Jaeger, F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in psychology*, 4.

Jin, L. and de Marneffe, M.-C. (2015). The overall markedness of discourse relations. In *EMNLP*.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329.

Kaiser, E. (2009). Effects of anaphoric dependencies and semantic representations on pronoun interpretation. In *Anaphora processing and applications*, pages 121–129. Springer.

Keenan, J., Baillet, S., and Brown, P. (1984). The effects of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior*, 23(2):115–126.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. CSLI publications Stanford.

Kehler, A., Kertz, L., Rohde, H., and Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, 25(1):1–44.

Kertz, L., Kehler, A., and Elman, J. (2006). Grammatical and coherence-based factors in pronoun interpretation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1605–1610.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, 95(2):163.

Kintsch, W. and Van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Knott, A. (1996). A data-driven methodology for motivating a set of coherence relations.

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(2):135–175.

Köhne, J. and Demberg, V. (2013). The time-course of processing discourse connectives. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

König, E. (1991). Concessive relations as the dual of causal relations. *Semantic Universals and Universal Semantics. Groningen-Amsterdam Studies in Semantics*, 12:190–209.

Koornneef, A. W. and Van Berkum, J. J. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, 54(4):445–465.

Kravtchenko, E. (2014). Predictability and syntactic production: Evidence from subject omission in russian. In *Proceedings of the 36th annual conference of the cognitive science society*.

Kuperberg, G., Paczynski, M., and Ditman, T. (2011). Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience*, 23(5):1230–1246.

Kurumada, C. and Jaeger, T. F. (2013). Communicatively efficient language production and case-marker omission in japanese. In *The 35th Annual Meeting of the Cognitive Science Society*, pages 858–863.

Lakoff, R. (1971). If's, and's and but's about conjunction.

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*, volume 71. Cambridge university press.

Lapata, M. and Lascarides, A. (2004). Inferring sentence-internal temporal relations. In *HLT-NAACL*, pages 153–160.

Lau, E. F., Holcomb, P. J., and Kuperberg, G. R. (2013). Dissociating n400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience*, 25(3):484–502.

Lea, R. B. and Mulligan, E. J. (2002). The effect of negation on deductive inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(2):303.

Lee, A., Prasad, R., Joshi, A., Dinesh, N., and Webber, B. (2006). Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax. In *Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories, Prague, Czech Republic, December*.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Levinson, S. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. The MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*.

Lewis, D. M. (2006). Discourse markers in english: a discourse-pragmatic view. *Approaches to discourse particles*, pages 43–60.

Lin, Z., Kan, M., and Ng, H. (2009). Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2010). A pdtb-styled end-to-end discourse parser. Technical report, Cambridge Univ Press.

Litman, D. J. (1996). Cue phrase classification using machine learning. *arXiv preprint cs/9609102*.

Longacre, R. E. (1996). *The grammar of discourse*. Springer Science & Business Media.

Louis, A., Joshi, A., Prasad, R., and Nenkova, A. (2010). Using entity features to classify implicit discourse relations. In *Proceedings of the 11th Annual Meeting of the Special*

Interest Group on Discourse and Dialogue, pages 59–62. Association for Computational Linguistics.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Mann, W. and Thompson, S. (1987). Rhetorical structure theory: A theory of text organization (no. isi/rs-87-190). marina del rey. *CA: Information Sciences Institute*.

Mann, W. and Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Marcu, D. (1997). The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–103. Association for Computational Linguistics.

Marcu, D. and Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375. Association for Computational Linguistics.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

McKeown, K. and Biran, O. (2013). Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 69–73. The Association for Computational Linguistics.

Millis, K., Golding, J., and Barker, G. (1995). Causal connectives increase inference generation. *Discourse Processes*, 20(1):29–49.

Millis, K. and Just, M. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language*.

Miltsakaki, E., Robaldo, L., Lee, A., and Joshi, A. (2008). Sense annotation in the penn discourse treebank. In *Computational Linguistics and Intelligent Text Processing*, pages 275–286. Springer.

Mitchell, J., Lapata, M., Demberg, V., and Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206.

Moser, M. and Moore, J. D. (1996). Toward a synthesis of two accounts of discourse structure. *Computational linguistics*, 22(3):409–419.

Murray, G., Taboada, M., and Renals, S. (2006). Prosodic correlates of rhetorical relations. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, pages 1–7. Association for Computational Linguistics.

Murray, J. (1995). Logical connectives and local coherence. *Sources of Coherence in Reading*, pages 107–125.

Murray, J. (1997). Connectives and narrative text: The role of continuity. *Memory and Cognition*, 25(2):227–236.

Nieuwland, M. S. and Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, 18(7):1098–1111.

Noordman, L. G. and Vonk, W. (1992). Readers' knowledge and the control of inferences in reading. *Language and Cognitive Processes*, 7(3-4):373–391.

Orenes, I., Beltrán, D., and Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language*, 74:36–45.

Otten, M. and Van Berkum, J. J. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, 45(6):464–496.

Park, J. and Cardie, C. (2012). Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112. Association for Computational Linguistics.

Patterson, G. and Kehler, A. (2013). Predicting the presence of discourse connectives. In *Proceedings of the conference on Empirical Methods in Natural Language Processing EMNLP*, pages 914–923.

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. (2008). Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*, 1(04):405–419.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2961–2968.

Prasad, R., Joshi, A., and Webber, B. (2010). Realization of discourse relations by other means: alternative lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1023–1031.

Pustejovsky, J., Havasi, C., Littman, J., Rumshisky, A., and Verhagen, M. (2006). Towards a generative lexical resource: The brandeis semantic ontology. In *Proceedings of the Fifth Language Resource and Evaluation Conference*, volume 7.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136.

Rohde, H. and Horton, W. S. (2014). Anticipatory looks reveal expectations about discourse relations. *Cognition*, 133(3):667–691.

Rohde, H. and Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8):912–927.

Rohde, H., Kehler, A., and Elman, J. L. (2006). Event structure and discourse coherence biases in pronoun interpretation. In *Proceedings of the 28th annual conference of the cognitive science society*, pages 697–702.

Rohde, H., Levy, R., and Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition*, 118(3):339–358.

Rutherford, A. T. and Xue, N. (2014). Discovering implicit discourse relations through brown cluster pair representation and coreference patterns.

Sanders, T. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes*, 24(1):119–147.

Sanders, T. (2005). Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning*, pages 105–114.

Sanders, T., Demberg, V., Evers-Vermeul, J., Hoek, J., Scholman, M., Asr, F. T., and Zufferey, S. (2016). Unifying dimensions in discourse relations: how various annotation schemes are related. In *3rd International Conference on Linguistic Psycholinguistic Approaches to Text Structuring (LPTS 2016)*.

Sanders, T., Spooren, W., and Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Sanders, T. J. and Noordman, L. G. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse processes*, 29(1):37–60.

Sayeed, A. B., Demberg, V., and Fischer, S. (2015). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the meeting of Association for Computational Linguistics*.

Schiffrin, D. (1988). *Discourse markers*. Number 5. Cambridge University Press.

Schiffrin, D. (2001). Discourse markers: language, meaning, and context. *The handbook of discourse analysis*, 1:54–75.

Schul, Y. (2011). Alive or not dead: Implications for framing from research on negations. *Perspectives on framing*, pages 157–176.

Segal, E., Duchan, J., and Scott, P. (1991). The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse Processes*, 14(1):27–54.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Sporleder, C. (2008). Lexical models to identify unmarked discourse relations: Does WordNet help? *Lexical-Semantic Resources in Automated Discourse Analysis*, page 20.

Staab, J. (2007). *Negation in context: Electrophysiological and behavioral investigations of negation effects in discourse processing*. ProQuest.

Stede, M. (2007). Rst revisited: Disentangling nuclearity. *2007):Subordinationversus coordinationin sentence and text–from a cross-linguistic perspective. Amsterdam.*

Stede, M. (2011). Discourse processing. *Synthesis Lectures on Human Language Technologies*, 4(3):1–165.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592.

Taboada, M. and Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, 4(2):249–281.

TextLink (2015). Structuring discourse in multilingual europe (textlink). [Online; accessed 27-September-2015].

Tily, H. and Piantadosi, S. (2009). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*.

Trabasso, T., Secco, T., and van den Broek, P. (1984). Causal cohesion and story coherence. *Learning and Comprehension of Text*, pages 83–111.

Trabasso, T., Secco, T., and Van Der Broek, P. (1982). Causal cohesion and story coherence. *Learning and comprehension of text*.

Trabasso, T. and Sperry, L. L. (1985). Causal relatedness and importance of story events. *Journal of Memory and language*, 24(5):595–611.

Traxler, M., Sanford, A., Aked, J., and Moxey, L. (1997). Processing causal and diagnostic statements in discourse. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):88.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443.

Van Petten, C. and Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and erp components. *International Journal of Psychophysiology*, 83(2):176–190.

Versley, Y. (2011a). Multilabel tagging of discourse relations in ambiguous temporal connectives. In *Proceedings de la 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*, pages 154–161.

Versley, Y. (2011b). Towards finer-grained tagging of discourse connectives. In *Proceedings of the Workshop Beyound Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*.

Versley, Y. (2013). Subgraph-based classification of explicit and implicit discourse relations. In *Proceedings of the International Conference on Computational Semantics, Potsdam, Germany*.

Wang, W., Su, J., and Tan, C. (2010). Kernel based discourse relation recognition with temporal ordering information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 710–719.

Wang, X., Li, S., Li, J., and Li, W. (2012). Implicit discourse relation recognition by selecting typical training examples. In *Proceedings of the COLING*, pages 2757–2772.

Webber, B. (2013). What excludes an alternative in coherence relations? In *Proceedings of the IWCS*.

Webber, B. and Joshi, A. (2012). Discourse structure and computation: past, present and future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54. Association for Computational Linguistics.

Webber, B., Knott, A., Stone, M., and Joshi, A. (1999). Discourse relations: A structural and presuppositional account using lexicalised tag. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 41–48. Association for Computational Linguistics.

Wellner, B., Pustejovsky, J., Havasi, C., Rumshisky, A., and Sauri, R. (2006). Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 117–125. Association for Computational Linguistics.

Wicha, N. Y., Moreno, E. M., and Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and

gender agreement in spanish sentence reading. *Journal of cognitive neuroscience*, 16(7):1272–1288.

Wilson, D. and Sperber, D. (2002). Relevance theory. *Handbook of pragmatics*.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Wlotko, E. W. and Federmeier, K. D. (2012). So that's what you meant! event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage*, 62(1):356–366.

Wolf, F., Gibson, E., Fisher, A., and Knight, M. (2005). The discourse graphbank: A database of texts annotated with coherence relations. *Linguistic Data Consortium*.

Xiang, M. and Kuperberg, G. (2014). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, (ahead-of-print):1–25.

Xu, X., Jiang, X., and Zhou, X. (2015). When a causal assumption is not satisfied by reality: differential brain responses to concessive and causal relations during sentence comprehension. *Language, Cognition and Neuroscience*, 30(6):704–715.

Xue, N., Ng, H. T., Pradhan, S., Bryant, R. P. C., and Rutherford, A. T. (2015). The conll-2015 shared task on shallow discourse parsing. *CoNLL 2015*, page 1.

Zhou, Z., Xu, Y., Niu, Z., Lan, M., Su, J., and Tan, C. (2010). Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514.

Zipf, G. K. (1949). Human behavior and the principle of least effort.

Zitoune, F. B. and Taboada, M. (2015). Mapping different rhetorical relation annotations: A proposal. *Lexical and Computational Semantics (* SEM 2015)*, page 147.

Appendix 1

	Estimate	Std. Error	t value	P-value
(Intercept)	0.7200	0.0183	39.34	0.0000
COMPARISON.Concession	0.0191	0.0913	0.21	0.8341
COMPARISON.Concession.contra-expectation	0.0942	0.0228	4.12	0.0000
COMPARISON.Concession.expectation	0.2079	0.0279	7.45	0.0000
COMPARISON.Contrast	-0.0591	0.0197	-3.01	0.0026
COMPARISON.Contrast.juxtaposition	-0.1017	0.0208	-4.88	0.0000
COMPARISON.Contrast.opposition	-0.0034	0.0267	-0.13	0.8997
COMPARISON.Pragmatic_concession	0.2800	0.1162	2.41	0.0159
COMPARISON.Pragmatic_contrast	0.1891	0.0769	2.46	0.0140
CONTINGENCY	-0.2200	0.1528	-1.44	0.1501
CONTINGENCY.Cause	-0.7200	0.4296	-1.68	0.0937
CONTINGENCY.Cause.reason	-0.3408	0.0196	-17.39	0.0000
CONTINGENCY.Cause.result	-0.4179	0.0203	-20.60	0.0000
CONTINGENCY.Condition	0.2800	0.3040	0.92	0.3571
CONTINGENCY.Condition.factual_past	0.2800	0.1442	1.94	0.0522
CONTINGENCY.Condition.factual_present	0.2800	0.0486	5.76	0.0000
CONTINGENCY.Condition.general	0.2800	0.0300	9.34	0.0000
CONTINGENCY.Condition.hypothetical	0.2787	0.0240	11.60	0.0000
CONTINGENCY.Condition.unreal_past	0.2800	0.0612	4.57	0.0000
CONTINGENCY.Condition.unreal_present	0.2800	0.0428	6.54	0.0000
CONTINGENCY.Pragmatic_cause.justification	-0.5821	0.0495	-11.75	0.0000
CONTINGENCY.Pragmatic_condition.implicit_assertion	0.2800	0.0665	4.21	0.0000
CONTINGENCY.Pragmatic_condition.relevance	0.2324	0.0954	2.44	0.0149
EXPANSION	-0.5149	0.0437	-11.78	0.0000
EXPANSION.Alternative	0.1848	0.0687	2.69	0.0072
EXPANSION.Alternative.chosen_alternative	-0.3109	0.0313	-9.94	0.0000
EXPANSION.Alternative.conjunctive	0.1133	0.0584	1.94	0.0521
EXPANSION.Alternative.disjunctive	0.2800	0.0403	6.95	0.0000
EXPANSION.Conjunction	-0.1217	0.0189	-6.44	0.0000
EXPANSION.Exception	0.1550	0.1089	1.42	0.1545
EXPANSION.Instantiation	-0.5460	0.0210	-26.00	0.0000
EXPANSION.List	-0.3339	0.0250	-13.34	0.0000
EXPANSION.Restatement	-0.6424	0.0336	-19.12	0.0000
EXPANSION.Restatement.equivalence	-0.6779	0.0313	-21.64	0.0000
EXPANSION.Restatement.generalization	-0.6427	0.0350	-18.36	0.0000
EXPANSION.Restatement.specification	-0.6764	0.0201	-33.59	0.0000
TEMPORAL	0.1891	0.1307	1.45	0.1480
TEMPORAL.Asynchronous	0.2800	0.2485	1.13	0.2598
TEMPORAL.Asynchronous.precedence	-0.0710	0.0214	-3.31	0.0009
TEMPORAL.Asynchronous.succession	0.1413	0.0224	6.30	0.0000
TEMPORAL.Synchrony	0.1612	0.0215	7.49	0.0000