# Coherence relations in discourse and cognition: Comparing approaches, annotations, and interpretations

Dissertation
zur Erlangung des akademischen Grades
eines Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes

vorgelegt von

Merel Cléo Johanna Scholman
aus Nieuwegein, die Niederlande

Saarbrücken, 2019

Dekan der Fakultät P: Prof. Dr. Heinrich Schlange-Schöningen

Erstberichterstatter: Vera Demberg

Zweitberichterstatter: Ted J.M. Sanders

Drittberichterstatter: Matthew Crocker

Tag der letzten Prüfungsleistung: 22 Februar 2019

## Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig verfasst habe, und dass es meine eigene Forschung beschreibt. Keine anderen als die angegebenen Quellen und Hilfsmittel sind verwendet.

## Declaration

I hereby declare that I composed this thesis entirely myself and that it describes my own research. I have not used any literature or materials other than the ones referred to in this thesis.

Merel C.J. Scholman
Saarbrücken
October 31, 2018

# Abstract

When readers comprehend a discourse, they do not merely interpret each clause or sentence separately; rather, they assign meaning to the text by creating semantic links between the clauses and sentences. These links are known as *coherence relations* (cf. Hobbs, 1979; Sanders, Spooren & Noordman, 1992). If readers are not able to construct such relations between the clauses and sentences of a text, they will fail to fully understand that text. Discourse coherence is therefore crucial to natural language comprehension in general.

Most frameworks that propose inventories of coherence relation types agree on the existence of certain coarse-grained relation types, such as causal relations (relations types belonging to the causal class include CAUSE or RESULT relations), and additive relations (e.g., CONJUNCTIONS or SPECIFICATIONS). However, researchers often disagree on which finer-grained relation types hold and, as a result, there is no uniform set of relations that the community has agreed on (Hovy & Maier, 1995).

Using a combination of corpus-based studies and off-line and on-line experimental methods, the studies reported in this dissertation examine distinctions between types of relations. The studies are based on the argument that coherence relations are cognitive entities, and distinctions of coherence relation types should therefore be validated using observations that speak to both the *descriptive adequacy* and the *cognitive plausibility* of the distinctions.

Various distinctions between relation types are investigated on several levels, corresponding to the central challenges of the thesis. First, the distinctions that are made in *approaches to coherence relations* are analysed by comparing the relational classes and assessing the theoretical correspondences between the proposals. An interlingua is developed that can be used to map relational labels from one approach to another, therefore improving the interoperability between the different approaches.

Second, practical correspondences between different approaches are studied by evaluating datasets containing *coherence relation annotations* from multiple approaches. A comparison of the annotations from different approaches on the same data corroborate the interlingua, but also reveal systematic patterns of discrepancies between the frameworks that are caused by different operationalizations.

Finally, in the experimental part of the dissertation, *readers' interpretations* are investigated to determine whether readers are able to distinguish between specific types of relations that cause the discrepancies between approaches. Results from off-line and online studies provide insight into readers' interpretations of multi-interpretable relations, individual differences in interpretations, anticipation of discourse structure, and distributional differences between languages on readers' processing of discourse.

In sum, the studies reported in this dissertation contribute to a more detailed understanding of which types of relations comprehenders construct and how these relations are inferred and processed.

# Zusammenfassung

Wenn Leser einen Diskurs verstehen, interpretieren sie nicht nur jeden Satz einzeln, sondern sie geben dem Text eine Bedeutung, indem sie semantische Verbindungen zwischen den Sätzen bzw. Teilsätzen herstellen. Diese Verbindungen sind bekannt als *Kohärenzrelationen* (vgl. Hobbs, 1979; Sanders, Spooren & Noordman, 1992). Wenn es einem Leser nicht gelingt, solche Relationen zwischen den Teilsätzen eines Textes herzustellen, wird er den Text nicht vollständig verstehen. Das Erkennen und Verstehen von Diskurskohärenz ist daher entscheidend für das natürliche Sprachverständnis.

Kohärenzrelationen bestehen zwischen mindestens zwei Textabschnitten, die als Segmente oder Argumente bezeichnet werden. Es wird allgemein angenommen, dass die Relationen zwischen den Argumente in eine feste, begrenzte Anzahl von Typen unterteilt werden können. Vorschläge, die bestimmte Typen von Kohärenzrelationen unterscheiden, werden als "Frameworks" bezeichnet. Die meisten Frameworks stimmen darin überein, dass es eine bestimmte Anzahl von groben Relationstypen gibt. Dazu gehören z.B. kausale Relationen (Beispiele von Relationstypen, die zur kausalen Klasse gehören, sind Cause oder Result Relationen), negative Relationen (z.B. Contrast oder Concession Relationen), additive Relationen (z.B. Conjunctions oder Lists), und zeitliche Relationen (z.B. Chronological oder Synchronous Relationen). Allerdings sind sich die Forscher oft nicht einig über die genaue Anzahl von feineren Relationstypen. Aufgrund dessen gibt es bis dato keine einheitliche Menge von Relationen, auf die sich die Gemeinschaft geeinigt hat (Hovy & Maier, 1995). Einige Vorschläge präsentieren mehr als 70 Typen von Relationen, wie von Carlson, Marcu & Okurowski (2003) entwickelt, andere präsentieren nur zwei Typen von Relationen (Grosz & Sidner, 1986).

Die in dieser Dissertation präsentierten Studien beschäftigen sich alle mit Unterschieden zwischen Relationstypen. Diese Unterschiede werden auf mehreren Ebenen untersucht. Die Unterscheidungen, die die einzelnen *Frameworks zu Kohärenzrelationen* machen werden analysiert, indem die Relationsklassen verglichen und die Übereinstimmungen zwischen den Frameworks bewertet werden. Ähnlichkeitenn zwischen verschiedenen Ansätzen werden untersucht, indem Datensätze ausgewertet werden, die *Annotationen von Kohärenzrelationen* von mehreren Ansätzen enthalten. Im experimentellen Teil der Arbeit werden *Leserinterpretationen* untersucht, um festzustellen, ob die Leser in der Lage sind, zwischen bestimmten Typen von Relationen zu unterscheiden, die Diskrepanzen zwischen den Ansätzen verursachen. Insgesamt liefern die hier präsentierten Studien ein detaillierteres Verständnis dafür, welche Typen von Relationen Leser bei der Sprachverarbeitung konstruieren und wie diese Relationen inferiert und verarbeitet werden.

Kapitel 3 gibt einen Überblick über verschiedene Frameworks, die Inventare von Relationsklassen vorgeschlagen haben. Insbesondere konzentriert sich diese Dissertation auf drei konkrete Frameworks: Penn Discourse Treebank (PDTB; Prasad et al.,

2008), Rhetorical Structure Theory Discourse Treebank (RST-DT; Carlson & Marcu, 2001), und Cognitive approach to Coherence Relations (CCR; Sanders et al., 1992). Die PDTB- und RST-DT-Frameworks wurden verwendet, um den größten verfügbaren englischen Korpus zu annotieren, und Varianten dieser Frameworks wurden verwendet, um Korpora in anderen Sprachen zu annotieren. Oft bestehen große Unterschiede zwischen den relationalen Inventaren. Dies erschwert es den Forschern oftmals, Daten zu verwenden, die nach dem anderen Framework annotiert wurden.

Aufbauend auf CCR wird eine neue Lösung – genannt Unifying Dimensions oder UniDim – vorgeschlagen, um die verschiedenen Ansätze miteinander in Beziehung zu setzen. Unifying Dimensions schlägt eine Interlingua vor, die die Relationslabels von einem Framework auf das andere abbilden kann und so die Übereinstimmung zwischen verschiedenen Ansätzen verbessert. Mit Hilfe der Unifying Dimensions sind Forscher in der Lage Daten zu verwenden die nach einem anderen Ansatz annotiert wurden.

Das Mapping von UniDim basiert auf den Definitionen und Beispielen, die in den Annotationsrichtlinien enthalten sind. Im Idealfall entsprechen die theoretischen Erläuterungen in den Richtlinien den tatsächlichen Annotationen in den Korpora, und die theoretische Abbildung würde daher auch der Abbildung in der Praxis entsprechen. Um dies zu untersuchen und die theoretische Kartierung zu validieren, haben wir Annotationen verschiedener Frameworks anhand der gleichen Daten verglichen, um sicherzustellen, dass die bereitgestellten Annotationen einander entsprechen.

Dieses Prozedere wurde für zwei Kombinationen von Frameworks durchgeführt: (i) wir haben einen Korpus von gesprochenen Daten sowohl mit PDTB als auch mit CCR annotiert (wie in Rehbein, Scholman & Demberg, 2016, beschrieben), was einen Vergleich der beiden Annotationsschichten ermöglichte; und (ii) wir haben die Annotationen von 385 Zeitungsartikeln des Wall Street Journal verglichen, die sowohl im PDTB als auch im RST-DT enthalten sind (wie in Demberg, Asr & Scholman, submitted, beschrieben). Kapitel 4 präsentiert beide Studien und ihre Ergebnisse.

Die wichtigsten Ergebnisse dieser Studien können wie folgt zusammengefasst werden: (i) die Unifying Dimensions konnten erfolgreich Übereinstimmungen zwischen den Frameworks vorhersagen, und (ii) es gab inhärente Muster von Meinungsverschiedenheiten in den Annotationen, die durch die Operationalisierungen der Frameworks verursacht wurden. Diese Cluster werden experimentell in den Studien untersucht, die in späteren Kapiteln vorgestellt werden.

Kapitel 5 bietet eine Reflexion darüber, wie relationale Unterschiede in Ansätze begründet werden können. Konkret werden zwei allgemeine Rechtfertigungsmaße diskutiert, die für linguistische Theorien verwendet wurden: *beschreibende Angemessenheit* (i.e., descriptive adequacy) und *kognitive Plausibilität* (i.e., cognitive plausibility). Viele Kohärenzrelationsansätze konzentrieren sich hauptsächlich auf die deskriptive Angemessenheit ihrer Inventare, d.h. das Inventar wird entwickelt, um alle Relationen in Texten zu beschreiben. Wir argumentieren, dass, um eine allgemeinere

Menge von Relationsstypen zu entwickeln, Unterscheidungen zwischen Kohärenzrelationen, dadurch gerechtfertigt werden können, dass nicht nur die Intuitionen von Experten (in Bezug auf die deskriptive Angemessenheit), sondern auch Erkenntnisse aus Akquisitions-, Produktions- und Verständnisstudien berücksichtigt werden. Mit anderen Worten, deskriptiv adäquate Ansätze können als Ausgangspunkt für die Entwicklung einer allgemeinen Theorie dienen: Sie können eine Bestandsaufnahme aller möglichen relationalen Konstrukte liefern, die dann anhand kognitiver Beweise validiert (verifiziert oder verfälscht) werden können. Theorien, die sich mit beiden Maßnahmen befassen, können unser Verständnis der mentalen Prozesse von Diskursdarstellungen, in der Sprachproduktion und im Sprachverständnis verbessern.

Es gibt bisher keine Quelle dazu welche Kriterien herangezogen werden können um zu beurteilen ob eine bestimmte relationale Unterscheidung oder ein Kohärenzrahmen kognitiv plausibel ist. Die Literatur sagt im Allgemeinen, dass Theorien auf empirischen Ergebnissen der Kognitionsforschung basieren müssen. Die spezifischen Arten von empirischen Ergebnissen oder kognitiver Forschung werden jedoch nicht näher ausgeführt. Kapitel 5 macht daher das Kriterium der kognitiven Plausibilität greifbarer, indem es detailliert beschreibt, welche Beweisquellen die Unterschiede zwischen kohärenten relationalen Labels und Klassen verifizieren oder falsifizieren können. Der kognitive Status von Kohärenzrelationen kann mit diesem Ansatz systematisch und schlüssig untersucht werden. Das Kriterium der kognitiven Plausibilität bildet die Motivation für alle weiteren Studien im weiteren Verlauf dieser Arbeit: Spezifische Unterschiede, über die sich die Frameworks nicht einig sind, werden bewertet, indem untersucht wird, ob naïve Leser diese Unterschiede vornehmen können und ob sie ihre Interpretationsprozesse beeinflussen.

Der zweite Teil dieser Dissertation beschreibt eine Reihe von experimentellen Studien, die auf Erkenntnissen aus vorherigen Kapiteln aufbauen. Die übergeordnete Frage für diese Kapitel lautet: Können Sprachbenutzer zwischen bestimmten Arten von Relationen unterscheiden, die Diskrepanzen zwischen den Ansätzen verursachen?

Kapitel 6 präsentiert eine Untersuchung des Designs und der Zuverlässigkeit einer Methode, die für die Verwendung in den Kapiteln 7 und 8 entwickelt wurde. Traditionelle Aufgaben der Diskursannotation gelten als kostspielig und zeitaufwendig, und die Zuverlässigkeit und Gültigkeit dieser Aufgaben werden in Frage gestellt. Hier wird daher eine neue Crowdsourcing-Methode entwickelt und evaluiert, bei der die Teilnehmer aufgefordert werden, aus einer vordefinierten Liste ein Bindewort auszuwählen, das die Verbindung zwischen den Segmenten einer Relation ausdrücken kann. Diese Methode ermöglicht es, die Diskursinterpretationen der Leser zu elizitieren, aus denen dann wiederum Diskursannotationen abgeleitet werden können. Wir haben auch den Einfluss des Kontextes auf die Zuverlässigkeit der Aufgabe untersucht, um mehr Einblick in die optimale Gestaltung einer solchen Aufgabe zu erhalten.

Die Ergebnisse der “Crowdsourced Connective Insertion Task” zeigten, dass die

Mehrheit der eingesetzten Bindewörter mit dem Original-Label konvergierte. Weiterhin zeigte die Verteilung der eingefügten Konnektive, dass oft mehrere Bedeutungen für eine einzelne Relation inferiert werden können. In Bezug auf die Anwesenheit von Kontext, zeigten die Ergebnisse keinen signifikanten Unterschied in der Verteilung der eingefügten Konnektive zwischen den Bedingungen insgesamt. Zusammenfassend deuten die Ergebnisse darauf hin, dass die neu entwickelte Crowdsourcing-Methode das Potenzial hat, als zuverlässige Alternative zu herkömmlichen Annotationsmethoden zu fungieren. Darüber hinaus liefern die Verteilungen der eingefügten Konnektive den Beweis, dass Relationen mehrere Interpretationen haben können, was wichtige Auswirkungen auf zukünftige Diskursannotationstudien haben könnte.

Kapitel 4 ergab, dass PDTB- und RST-DT-Annotatoren oft über die Annotation der beiden kohärenzrelationalen Typen Beispielen und Spezifikationen anderer Meinung sind; insbesondere gibt es Unterschiede bezüglich der Interpretation dieser beiden Typen als ideelle (additive) oder argumentative (pragmatische kausale) Kohärenzrelationen. Das Kapitel 7 untersuchte daher, wie naïve-Leser diese Relationen interpretieren, indem sie die in Kapitel 6 vorgestellte Crowdsourced-Methode verwendeten. Die Ergebnisse zeigten, dass diese Relationen tatsächlich zwei Funktionen haben können: Sie können sowohl zur Veranschaulichung / Spezifizierung einer Situation als auch als Argument für eine Behauptung verwendet werden. Diese Ergebnisse deuten darauf hin, dass Beispiele und Spezifikationen mehrere, gleichzeitige Interpretationen haben können.

Das Kapitel 8 untersucht weiter, wie die Leser Relationen mit mehreren möglichen Interpretationen interpretieren. Konkret haben wir untersucht, ob die Leser Präferenzen für eine bestimmte Interpretation haben und ob es individuelle Unterschiede in diesen Präferenzen gibt. Die Crowdsourced-Methode wurde in einem Messwiederholungsdesign verwendet.

Die Ergebnisse zeigten, dass die Teilnehmer konsistente Präferenzen bei der Interpretation von Kohärenzrelationenn hatten und dass sie sich in diesen Präferenzen voneinander unterschieden. Darüber hinaus schienen sich die Teilnehmer in ihrer "standardmäßigen" Verarbeitungstiefe zu unterscheiden, was einige der Unterschiede in der Interpretation erklärt. Doch selbst wenn die Leser eine Aufgabe erfüllten, die von ihnen eine tiefe Verarbeitung verlangte, blieben einige Unterschiede in der Interpretation bestehen. Die Ergebnisse einer verbalen Arbeitsgedächtnisaufgabe zeigten, dass die in der vorherigen Studie gefundenen individuellen Unterschiede nicht durch Unterschiede im Arbeitsgedächtnis erklärt werden konnten. Die Ergebnisse der aktuellen Studie zeigten daher, dass die Leser individuelle Variabilität in ihren Interpretationspräferenzen von ambigen Relationen aufweisen, in Abhängigkeit davon wie tief sie Text verarbeiten. Theorien und Experimente zur Diskursinterpretation sollten deswegen die Unterschiede in den Interpretationspräferenzen und der Verarbeitungstiefe berücksichtigen.

In den letzten beiden Kapiteln konzentrieren wir uns auf die Interaktion zwischen

Konnektiven und dem Inhalt von Segmenten. Konnektive sind wichtige Signale für Kohärenzrelationen: Die Forschung hat gezeigt, dass sie als Verarbeitungsanweisungen funktionieren, indem sie signalisieren, wie die Segmente miteinander verbunden werden sollen. Es wird davon ausgegangen, dass die Leser sowohl das Konnektiv als auch den Inhalt der Segmente verarbeiten, um eine Relation herzuleiten. Das PDTB–RST-DT-Mapping in Kapitel 4 zeigt jedoch, dass sich Leser mehr auf Konnektiven als auf den Inhalt der Argumente verlassen, um eine Relation abzuleiten: In Abwesenheit eines expliziten Konnektives zeigten die Annotatoren wenig Übereinstimmung über die Art der Relation, die sie inferierten. Der letzte Teil der Arbeit untersucht, ob die Leser sowohl die Konnektive als auch den Inhalt der Segmente nutzen, um Kohärenzrelationen abzuleiten.

Die Cue-Phrase *"On the one hand"* (Einerseits; OT1H) schafft die Erwartung einer kontrastiven Relation. Kapitel 9 testet ob diese Erwartung durch ein kontrastierendes Bindeglied nach OT1H erfüllt werden kann, oder ob die Leser in der Lage sind, den Inhalt der Segmente zu berücksichtigen und eine kontrastive Relation mit dem entsprechenden Segment herzustellen – auch wenn dies nicht das erste kontrastive Segment nach OT1H ist. Wir verwendeten kurze Passagen mit *"On the one hand"* und *"On the other hand"* (Andererseits; OTOH) sowie einen dazwischen liegenden kontrastierenden Satz mit *but* (aber). Es gab zwei Versionen dieses intervenierenden kontrastierenden Satzes: Ein lokal kontrastierender Satz stellte keinen angemessenen Kontrast zum OT1H-Satz her; ein global kontrastierender intervenierender Satz tat es. Wenn die Leser den Inhalt der Segmente berücksichtigen, sollten sie auf den Unterschied zwischen lokal und global kontrastierendem Satz achten.

Drei Experimente mit Offline- und Online-Messungen zeigten, dass die Befragten mehr als nur eine passende Form erwarteten (d.h. *On the other hand*): Leser behielten ihre Erwartung an einen bevorstehenden Kontrast über das dazwischenliegende Material bei, auch wenn der eingebettete Bestandteil selbst Kontrast enthielt. Darüber hinaus führte ein nachfolgender Kontrast, der mit *On the other hand* markiert war, bei den Lesern zu Verarbeitungsschwierigkeiten, vor allem wenn zuvor ein anderer Textabschnitt Informationen enthielt, die einen angemessenen Kontrast zu *On the one hand* herstellten.

Die Eye-Tracking-Studie in Kapitel 9 untersuchte auch, ob das Vorhandensein von OT1H die Verarbeitung von OTOH erleichtert. Die Ergebnisse zeigten aber keinen solchen Effekt. In diesem Kapitel wurden zwei zentrale Hypothesen getestet. Eine erste Hypothese war, dass ein möglicher facilitativer Effekt von OT1H reduziert wurde, weil die Erwartungen an OTOH insgesamt während der gesamten Studie gestiegen sind, was auf eine große Anzahl von OTOH-Items zurückzuführen ist. Deshalb wurde eine englische Eye-Tracking-Studie mit weniger OTOH-Items durchgeführt. Die Ergebnisse zeigten eine unterstützende Wirkung von OT1H auf die Interpretation von OTOH, was darauf hindeutet, dass die Häufigkeit von OTOH in der früheren Studie tatsächlich einen möglichen Effekt verbarg.

Eine zweite Hypothese war, dass OT1H im Englischen keine starke unterstützende

Wirkung hat, da OTOH in natürlicher Sprache oft verarbeitet wird ohne dass vorher OT1H erschienen ist. Da OT1H mit OTOH im Niederländischen häufiger auftritt als im Englischen, wurde untersucht, ob die unterstützende Wirkung von OT1H im Niederländischen größer ist. Wieder führten wir eine Eye-Tracking-Studie durch, die tatsächlich eine stärkere unterstützende Wirkung von OT1H im Niederländischen im Vergleich zum Englischen ergab. Dies deutet darauf hin, dass die Verarbeitung von durch sprachspezifische Faktoren wie Verteilungsmerkmale beeinflusst wird, sodass sich daher sprachübergreifend unterschiedliche Präferenzmuster zeigen können.

Insgesamt werden in dieser Dissertation verschiedene Studien vorgestellt die die Unterscheidung zwischen den Typen von Kohärenzrelationen untersuchen, indem Ansätze, Annotationen und Interpretationen miteinander verglichen werden. Diese Studien liefern Einblicke in die Ähnlichkeiten und Unterschiede zwischen den Inventaren, die für Unterscheidungen zwischen den Kohärenzrelationen vorgeschlagen wurden,, sowie in die Operationalisierungen dieser Ansätze. Insgesamt wird gezeigt wie sich diese Faktoren auf die resultierenden Annotationen auswirken können. Durch die Untersuchung, wie verschiedene Arten von Relationen von naïve-Lesern interpretiert und verarbeitet werden, lieferte diese Dissertation neue Einblicke in die Prozesse des Diskursverständnisses.

# Acknowledgment

The list of people to whom I am indebted is long. First, I am very grateful to Vera Demberg and Ted Sanders. Vera, thank you for taking me on and giving me the freedom to pursue my interests. You have created a lab that combines excellent working conditions (also thinking of the *Rest* book here) and a positively challenging research environment, and I definitely benefited from this. Thank you for encouraging me to grow and trusting me with certain responsibilities that contributed towards this growth. Most importantly, thank you for listening, and for being flexible and understanding when I needed it the most.

Ted, thank you for your support and guidance, and for introducing me to the wonderful world of discourse. Over the years, you have held several roles in my academic career: a guest lecturer, my professor, my “boss”, then theses supervisor, and now the second reader for this dissertation. Above all, you have been a mentor to me in various facets of my life, and for this I cannot thank you enough. I would not have made it without your encouragement and continuous belief in me.

I have also been fortunate enough to get to collaborate closely with other inspiring researchers. Jacqueline Evers-Vermeul, we’ve managed to work together on quite a few projects. Thank you for your encouragement these past years, and for many enjoyable lunches, tea breaks, and dinners. Hannah Rohde, thanks for hosting me in Edinburgh, and for introducing me to R. Kate Cain, thanks for hosting me in Lancaster, and for being so generous in rallying lab members to collect data (and thanks to Liam Blything, Gill Francey, and Nicola Currie for their help in collecting the data!). Ines Rehbein, thank you for the summer of annotation discussions when I first arrived in Saarbruecken. Thanks also to Torgrim Solstad and Elsi Kaiser for invaluable discussions that we had during your lab visits.

I consider myself particularly lucky that I conducted my PhD research at the same time as Jet Hoek (albeit in different countries). Jet, from the very beginning of my PhD, you have been there for me, helping me construct items for experiments, examples for presentations, questions for my defense, and even logos for my wedding. Many ideas in this dissertation benefited from being run by you during early stages. What’s more, I personally benefited from being able to travel with you so often, which made the past few years fly by. I couldn’t have wished for a better Peppi to my Kokki.

I also want to thank my (ex-)colleagues in Saarbrücken: Alessandra Zarcone, David Howcroft, Elisabeth Rabs, Fatemeh Torabi Asr, Frances Yung, Iona Gessinger, Jorrig Vogels, Katja Häuser, Katja Kravtchenko, Michael Roth, Tony Hong, and Wei Shi. I have enjoyed our lively discussions and game nights. I am also grateful to my fellow Utrecht PhDs, who made me feel part of their discourse family: Andrea Santana C., Nina Sangers, Suzanne Kleijn, and Yipu Wei.

I would like to give special thanks to those offering to lend me a helping hand when mine would no longer work: Gabriele Reibold, thank you for your help with a (rather large) variety of issues, and for our talks about Nigerian culture. Romée van

Erning, thanks for being so willing to jump in, and for the time we spent working together before as well. Vera, Ted, Jacq, Jet, Katja H, my family, and many others: thanks for listening to my woes and helping me figure out solutions.

Two networks have supported my research. First, I would like to thank the center that funded my research, the Collaborative Research Center SFB 1102 on "Information Density and Linguistic Encoding." In particular, I would like to thank three professors who are part of the SFB, namely Matt Crocker, Elke Teich, and Ingo Reich, for agreeing to be part of my committee and for the valuable discussions we had. I would also like to thank the COST network TextLink (led by Liesbeth Degand) and all of its participants. The network gave me the opportunity to meet and discuss with many experts in the discourse field, from which the ideas in this dissertation benefited. A special thanks goes to Sandrine Zufferey for her advice and for the pleasant conversations that we have had.

There are many people outside of my academic environment to whom I am indebted as well, for keeping me happy, healthy, and sane. To my friends, especially those who have crossed borders and sometimes even oceans to spend time with me and celebrate with me: thank you for sticking with me. Aga, Branko, Flo, Frank, Jessie, Lili, Lizzy, Marlies, Rosan, and Rosanna: despite being even further away now than I have been, I hope we will continue to be there for each other and celebrate various life events together.

To my parents: thank you for your support, for taking me in whenever I needed care and comfort, and for respecting all of my choices and plans, no matter how crazy or difficult they might seem. I should also mention: thank you for all of your invaluable infrastructural support.

Thari, thank you for listening, and for giving me peace and another place to call home. Most importantly, thank you for "providing" me with three beautiful godchildren. Noah, Toby and Lilou: you have given me so much love, strength, much needed distraction, and heartache (because I miss you dearly).

Finally, Nnamdi: there is too much to thank you for. Thank you for your cheerfulness, encouragement, and understanding. Thank you for indulging many of my whims and wishes. Thank you for being there. Thank you for waiting. You know I dey mad over you. *Ihunanya m bu nke gi mgbe nile.*

Lagos, March 2019

# Contents

# Chapter 1

# Introduction

When readers comprehend a discourse, they do not merely interpret each clause or sentence separately; rather, they assign meaning to the text by creating semantic links between the clauses and sentences. These links are known as *coherence relations* (cf. Hobbs, 1979; Sanders et al., 1992). To illustrate the notion of coherence, consider the following example, which was featured as a Buzzfeed news article headline:

(1) Lil Wayne hospitalized after private plane makes emergency landing.

This sentence expresses two events that are linked in a temporal relationship, marked by the connective *after*: there was an emergency landing, after which Lil Wayne was hospitalized. Even though the connective *after* indicates a temporal relation, comprehenders are likely to interpret these events as being in a causal relationship: Lil Wayne was hospitalized *because* his private plane made an emergency landing.[1] Hence, comprehenders assign more meaning to this sentence than that conveyed by the individual clauses alone.

If readers are not able to construct such relations between the clauses and sentences of a text, they will fail to fully understand that text. Discourse coherence is therefore crucial to natural language comprehension in general. A more detailed understanding of which types of relations comprehenders construct and how these relations are inferred and processed would contribute to the development of a full-coverage theory of discourse coherence. This thesis aims to achieve a better understanding of the distinctions between coherence relations by investigating three central challenges:

(i) **The differences between various theories, and between the types of coherence relations they propose**
Various frameworks have proposed inventories of coherence relations. These

[1] In reality, Lil Wayne needed medical attention, which caused the emergency landing. Example 1 corresponds to the original Buzzfeed headline, it was later changed to "Lil Wayne hospitalized after having 2 seizures on private plane," likely in order to address the ambiguity of the relation.

frameworks have been used to analyse coherence by assigning labels to the relations (the process of annotation) and collecting these annotated instances in corpora. The differences between the proposed inventories make it difficult to draw comparisons across the corpora annotated in the various frameworks. In this thesis, we investigate the differences and similarities between these proposals in order to increase the interoperability between them. Specifically, we look at the types of coherence relations that they distinguish, and determine how these types relate to each other (that is, we're looking at "relations among relations"). For example, the label CIRCUMSTANCE is distinguished in one framework, but not another. By being able to identify equivalent labels in other frameworks, researchers working in different paradigmata could make use of all the annotated data that is available.

(ii) **Patterns of discrepancies between annotations**

The comparison between annotations stemming from various frameworks provides more insight into the similarities and differences between these proposals. In the current thesis, these annotations are compared, thereby revealing distinct patterns of discrepancies between frameworks. An analysis of these patterns indicates that they are caused by systematic differences in the operationalizations of frameworks. They are therefore theoretical in nature. We argue that such discrepancies can be investigated by studying cognitive evidence in order to determine the validity of the coherence relational categories. One example of a systematic type of discrepancy is those relations that one framework annotates as SPECIFICATION, whereas another framework annotates these as EVIDENCE or EXPLANATION-ARGUMENTATIVE. Passage 2 illustrates this.

(2) The two executives could hardly be more different. Mr. Roman comes across as a low-key executive; Mr. Phillips has a flashier personality. wsj_1317

The second discourse segment of this example can be interpreted as specifying how the two executives are different, or as providing evidence for the claim that the two executives are very different. By investigating such patterns of discrepancies, we can gain a better understanding of how frameworks relate to each other and how they differ from each other. We can also gain more insight into which types of relations often convey multiple meanings.

(iii) **Interpretations of ambiguous relations**

Studies investigating readers' interpretations of coherence relations provide a crucial source of evidence for the cognitive plausibility of relational distinctions. In the current thesis, various methodologies are used to study these interpretations: connective insertion studies reveal distributions of readers' interpretations, story continuation studies reveal readers' interpretation of and sensitivity

> to relational signals, and eye-tracking studies provide insight into how readers process relations. Two types of "ambiguities" are investigated: relations that can have multiple interpretations (such as PDTB's SPECIFICATION relations), and relations that have intervening material which can be interpreted as relational segments, as in Example 3.
>
> (3) John is considering a job offer at the zoo.
> On the one hand, he really needs the money, because he should start paying off his student loans.
> But the loans could be deferred for a few more months.
> On the other hand, he hates the idea of cleaning out panda cages.
>
> The intervening sentence marked by *But* can be interpreted as satisfying the expectation for a contrast set up by *On the one hand.* However, the *On the other hand* actually satisfies this expectation, and so interpreting the intervening sentence as the counterpart of *On the one hand* would lead to a false interpretation of the relation. This thesis investigates whether readers can successfully construct relations in such situations.

Answers to questions that correspond to these challenges will contribute to the development of a general theory of coherence, the design of annotation and experimental studies, and the improvement of the accuracy of relational classifiers. The next section presents the research goals that are investigated in this dissertation, followed by the contributions. The chapter concludes with an overview of all chapters in this dissertation.

## 1.1 Research goals

Various proposals (also referred to as frameworks) have been put forward to describe a standard set of coherence relations. Increased interoperability between these frameworks and corpora would benefit the community for a number of reasons. For example, interoperability of corpora is useful for researchers working on automatic coherence relation labelling: interoperable corpora would allow researchers to train their models on multiple corpora, rather than just those corresponding to a specific framework. This would likely improve the performance of these tools.

**Goal 1.** The first goal of the thesis is therefore **to improve interoperability between frameworks by creating an intermediate coherence relation representation, referred to as the Unifying Dimensions approach**. This approach can be used to translate relational labels from one framework into those of another framework. Moreover, the representation improves our understanding of the features defining different types of relations, and allows us to pinpoint the exact differences and similarities between different types of relations. The approach addressed central challenge (i): identifying differences and similarities between coherence relation frameworks and the relational categories that they propose.

The approach was based on a theoretical mapping, which was developed using the definitions and examples provided in the annotation guidelines to determine the mappings between the frameworks. Ideally, the theoretical explanations provided in the guidelines correspond to the actual annotations in the corpora, and the theoretical mapping would therefore also correspond to the mapping and practice. However, this would need to be investigated.

**Goal 2.** The second goal of the thesis is **to validate the theoretical mapping and investigate the compatibility of annotations corresponding to different frameworks**. This was done by comparing annotations of various frameworks on the same data, to ensure that the provided annotations correspond to each other. The results mainly corroborate the theoretical mapping, but also reveal that agreement between these frameworks is relatively low. This effort addressed two central challenges: (i) to further investigate differences and similarities between frameworks, and (ii) to investigate patterns of discrepancies between annotations – that is, cases where annotators from one framework disagree with annotators from another framework. The discrepancies that are identified will function as input for later investigations in this thesis.

Given that a large number of different frameworks proposing different relational distinctions exists, and that the agreement between frameworks appears to be relatively low, we need to reevaluate how relational distinctions can be justified.

**Goal 3.** The third goal of the thesis is **to argue for more consistent methods of providing justification for relational distinctions, namely by providing evidence for the descriptive adequacy as well as the cognitive plausibility of theories**. Crucially, it is argued that cognitively plausible evidence for relational distinctions strengthens theories and provides justification that these theories model the actual constructs that are present in our linguistic system. A method is proposed that details what sources of evidence should be considered, and what counts as enough evidence or counter-evidence for the validity of the distinctions proposed in theories. This addresses central challenge (i): the relational categories that are distinguished in theories and frameworks are evaluated by considering evidence and counter-evidence.

One of the possible sources of evidence that is proposed to account for the cognitive plausibility of theories stems from experiments investigating interpretations of naïve readers (that is, readers who are not discourse analysts). Crowdsourcing provides an ideal platform to elicit discourse interpretations from naïve readers. However, a new method is needed that can crowdsource naïve readers' interpretations regarding which coherence relation holds between two segments of text.

**Goal 4.** A fourth goal is **to develop a new methodology that can be used to crowdsource interpretations of coherence relations**. This connective insertion method can be used for various research goals. In the current thesis, it is used to evaluate the discrepancies found in the mapping studies. However, the methodology also has the potential of becoming a useful tool to elicit coherence relation annotations in order to create a discourse annotated-corpus. This alternative method for annotating coherence relations can address the lack of training data for automatic coherence relation classifiers. The development of this method contributes to central challenges (ii) and (iii): the methodology is used to investigate patterns of discrepancies between annotations, and to elicit interpretations of ambiguous relations.

**Goal 5.** The fifth goal of the thesis is **to investigate the cause of a specific discrepancy between frameworks (and their corresponding annotations) using distributions of interpretations by naïve readers**. Additionally, this investigation addresses the *multi-level thesis* by providing evidence that comprehenders can infer relations on different levels of discourse. Individual differences in interpretation preferences are considered, thereby providing novel insight into the effect of individual variability on discourse interpretations. These studies address central challenge (ii) and (iii).

It is generally assumed that comprehenders make use of the connectives and other relational signals present in the discourse, as well as the contents of the segments in order to infer the coherence relations in a text. However, the evaluation of the compatibility between annotations revealed that annotators strongly rely on the connective in order to infer a relation. In the absence of a connective, the agreement between annotators on the interpretations of relations is much lower. The final chapters of this dissertation therefore focus on the nature of coherence inferences.

**Goal 6.** The final goal of the thesis is therefore **to investigate whether readers make use of both the connective and the content of the segments when processing discourse**. Furthermore, the effect of distributional differences of connectives in different languages on comprehenders' processing is investigated, in order to explore whether discourse expectations are modulated by language-specific factors. These experiments address central challenge (iii).

## 1.2 Contributions of the research

The main contributions of the work that is reported in this dissertation include the following:

- **Unifying Dimensions approach**
  The Unifying Dimensions approach – an intermediate coherence relation representation – is proposed, which can be used to translate relational labels from one framework into those of another framework. The approach improves interoperability between different coherence frameworks and the corpora annotated using these frameworks. An evaluation done on annotated data supports the mapping and highlights areas for future research.
  This work corresponds to research goals 1 and 2, and is described in Chapters 3

and 4 (joint work with Ted Sanders, Vera Demberg, Jacqueline Evers-Vermeul, Jet Hoek, Ines Rehbein, Fatemeh Torabi Asr and Sandrine Zufferey; reported in Demberg, Asr & Scholman, submitted; Rehbein, Scholman & Demberg, 2016; Sanders, Demberg, Hoek, Scholman, Torabi Asr, Zufferey & Evers-Vermeul, 2018).

- **Tangible criteria for cognitive plausibility**
  Tangible criteria for the justification of coherence relational categories are defined. We argue for a consistent method of justification based on evidence of the descriptive adequacy and cognitive plausibility of the distinctions. Such a method can contribute to the development of a general theory of coherence by providing a means of validating proposals.
  This contribution corresponds to research goal 3 and is reported in Chapter 5.

- **Crowdsourced interpretation elicitation method**
  A novel method is developed that can be used to elicit discourse interpretations from naïve readers. The method also has the potential of becoming a useful tool to elicit coherence relation annotations in order to create a discourse-annotated corpus.
  The method corresponds to research goal 4 and is described in Chapter 6 (joint work with Vera Demberg; reported in Scholman & Demberg, 2017a).

- **Evidence for the multi-level thesis and interpretation biases**
  Evidence is provided for the multi-level thesis by showing that comprehenders can infer relations on different levels of discourse. Additionally, evidence is found that comprehenders display individual differences in discourse interpretations: readers have consistent biases for interpreting one level over the other, and they exhibit individual variation in these biases.
  The studies correspond to research goal 5 and are described in Chapters 7 and 8 (joint work with Vera Demberg and Ted Sanders; reported in Scholman & Demberg, 2017b).

- **Evidence for sensitivity to discourse structure and cross-linguistic differences**
  Evidence is found that readers generate fine-grained expectations of upcoming coherence relations based on both connectives and the content of the segments. The importance of cross-linguistic research is highlighted by showing that discourse expectations are modulated by language-specific factors such as distributional differences.
  The experiments correspond to research goal 6 and are described in Chapters 9 and 10 (joint work with Vera Demberg and Hannah Rohde; reported in Scholman, Rohde & Demberg, 2017).

The implications of these studies are discussed in the final chapter, where several research directions are proposed for future work on coherence relation annotation projects as well as discourse interpretation and processing experiments.

## 1.3 Overview of the dissertation

The studies reported in this dissertation all examine the phenomenon of coherence relations; specifically, distinctions between types of relations. These distinctions are investigated on several levels, corresponding to the central challenges of the thesis: the distinctions that are made in *approaches to coherence relations* are analysed by comparing the relational classes and assessing the correspondences between the proposals. Correspondences in annotations from different approaches are studied by evaluating datasets containing *coherence relation annotations* from multiple approaches. In the experimental part of this thesis, *readers' interpretations* are investigated to determine whether readers are able to distinguish between specific types of relations that cause discrepancies between approaches.

**Chapter 2** provides the relevant background for the remaining chapters in the thesis. Specifically, it elaborates on the cognitive status of coherence relations, the segments between which a relation holds, which types of relations are distinguished, and how comprehenders construct coherence relations.

**Chapter 3** discusses several approaches that have put forward inventories of relational classes. These approaches are particularly relevant to the first chapters in this dissertation, and the description therefore provides the necessary background information. Chapter 3 also discusses attempts that have been made to unify the inventories of different approaches. Finally, a new solution – referred to as Unifying Dimensions – is proposed for relating the different approaches to each other. The Unifying Dimensions solution addresses the first research goal by proposing an interlingua that can map relation labels from one approach to another, therefore improving the interoperability between different approaches. Using the Unifying Dimensions, researchers that work with one specific approach are able to make use of data annotated according to another approach, despite differences in the relational labels and categories.

**Chapter 4** puts the proposed solution to the test by evaluating the mapping using actual annotations. This effort addresses the second research goal. Annotations from different approaches on the same data were compared, to investigate whether the provided annotations show the correspondences that were predicted by the Unifying Dimensions approach. Two investigations are reported in this chapter. The two main findings of these efforts were: (i) the Unifying Dimensions approach was successful in predicting the correspondences between the annotations of different frameworks, and (ii) there were inherent clusters of disagreements in the annotations that were caused by the operationalizations of the frameworks. These clusters are examined experimentally in the experiments reported in later chapters.

**Chapter 5** provides a reflection on how relational distinctions in proposals can be justified, thereby addressing the third research goal. The criterion of cognitive plausibility is concretized by nominating various sources of evidence that can support or justify particular distinctions. Crucially, it is argued that the experts' intuitions and the linguistic evidence should be supplemented with cognitive evidence stemming from experimental studies. This idea forms the motivation for the studies in the next chapters: specific distinctions or assumptions that frameworks disagree on are evaluated by examining whether readers infer these relations and can make these distinctions.

The second part of this dissertation describes a series of experimental studies that follow up on findings from earlier chapters. The overarching question for these chapters is: are language users able to distinguish between specific types of relations that cause discrepancies between approaches?

**Chapter 6** presents an investigation of the design and reliability of a methodology that is developed for use in Chapters 7 and 8. This method corresponds to the fourth research goal. Specifically, we designed a crowdsourcing method that can be used to obtain naïve readers' interpretations of relations. Participants were presented with two text segments and asked to insert a connective (from a predefined list) that best expresses the relation between the segments. From these insertions, we were able to derive a distribution of coherence relation labels for every item. The results showed that naïve readers can perform this task accurately (compared to experts annotators), even in the absence of linguistic context. The chapter ends with a discussion on the possibilities of using this method to annotate coherence relations.

**Chapter 7** explores one of the discrepancies identified in Chapter 4, thereby addressing the fifth research goal. The mapping studies showed that relations annotated in one approach as INSTANTIATION or SPECIFICATION (i.e., additive relations) were often annotated in another approach as EVIDENCE or EXPLANATION-ARGUMENTATIVE (i.e., causal relations). Using the crowdsourced connected insertion method, we investigated how naïve readers interpret these relations. The results showed that both readings can be inferred for the items, indicating that both annotations were justified. It is argued that these relations are multi-interpretable and that these findings support the multi-level theory. The chapter ends with a discussion of the implications of these results for this theory.

**Chapter 8** also addresses the fifth research goal by further investigating readers' interpretations of relations with multiple possible interpretations. Specifically, we investigated whether there were individual differences in readers' preferences for inferring one sense over another, and whether there was individual variability in these preferences. The connective insertion method was used in a repeated measures design. The results showed that comprehenders did indeed have a bias for inferring one sense over the other, and this bias was variable: some participants had a preference for inferring the causal sense, whereas others more often inferred the additive sense.

In a follow-up experiment, we found that participants were more likely to infer the additive sense when engaging in deeper processing. However, even when comprehenders performed a task that required them to process deeply, some differences in interpretations remained. The chapter ends with a discussion on the importance of accounting for differences in interpretation biases and depth of processing in theories and experiments on discourse interpretations.

In the final two chapters, we turn our focus to the interaction between connectives and the content of segments. Both chapters address the sixth and final research goal. Connectives are important signals of coherence relations: research has shown that they function as processing instructions during reading, by signalling to comprehenders how to link the segments. In other words, it is assumed that readers process both the connective and the content of the segments to infer a relation. However, the mappings in Chapter 4 indicate that comprehenders (or at least annotators) rely heavily on connectives, rather than the content of the segments, to infer a relation: in absence of an explicit connective, annotators show little agreement on the type of relation they infer. The final part of the thesis examines whether readers make use of both the connectives and the content of the segments to infer coherence relations.

**Chapter 9** used the pair *On the one hand / On the other hand* (OT1H/OTOH) to investigate whether comprehenders were sensitive to discourse structure, and whether they made use of connectives to predict upcoming coherence relations across longer passages. We used short passages containing OT1H/OTOH as well as an intervening contrastive sentence marked by *but*. Different versions of this intervening sentence were created: a locally contrastive sentence did not establish an appropriate contrast with the OT1H sentence; a globally contrastive intervening sentence did. If comprehenders relied solely on connectives, they should not be sensitive to the difference between locally and globally contrastive sentence. Three experiments using offline and online measures showed that comprehenders did take into account the content of an intervening contrastive sentence, and they therefore dispreferred a subsequent contrast marked with OTOH when the intervening content had established an appropriate contrast with OT1H. Furthermore, comprehenders maintained their expectation for an upcoming contrast across intervening material, doing so even if the intervening sentence itself contained contrast. The results are taken to support expectation-driven models of processing in which comprehenders posit and maintain fine-grained representations of discourse structure.

The results of the eye-tracking study in Chapter 9 failed to show that the presence of *On the one hand* facilitated the processing of *On the other hand*. **Chapter 10** follows up on this finding by testing several hypotheses. In particular, we investigated whether the lack of a facilitative effect could be explained by rapid expectation adaptation throughout the experiment: repeated exposure to the a priori unexpected structure of OTOH could have reduced its processing disadvantage in items where OT1H was absent. We tested this hypothesis in an eye-tracking study in which fewer items with these markers were included. A second hypothesis for the lack of an effect is related to the distribution of the pair of markers: OTOH occurs without OT1H relatively frequently in natural language. We therefore conducted a similar eye-tracking study in Dutch, using the Dutch equivalents of OT1H/OTOH. In Dutch, OTOH is foreshadowed by OT1H more frequently. The results suggested that both hypotheses hold: the English eye-tracking study now revealed a facilitative effect of OT1H on OTOH, and the Dutch eye-tracking study revealed an even stronger facilitative effect of OT1H in Dutch compared to English. The findings therefore suggest that the effect of discourse markers on comprehenders' processing is modulated by language specific factors and, as a result, can differ cross-linguistically.

## 1.4 Relevant publications

The following publications report on parts of the research described in this dissertation:

- Demberg, V., Asr, F.T. and Scholman, M.C.J. (submitted). How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Submitted to Dialogue & Discourse.*

- Rehbein, I., Scholman, M.C.J., and Demberg, V. (2016). Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 16)* (pp 23–28). Portoroz, Slovenia.

- Sanders, T.J.M., Demberg, V., Hoek, J., Scholman, M.C.J., Asr, F.T., Zufferey, S., and Evers-Vermeul, J. (2018). Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory, ahead of print.* DOI: https://doi.org/10.1515/cllt-2016-0078

- Scholman, M.C.J. and Demberg, V. (2017). Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. *Proceedings of the 11th Linguistic Annotation Workshop* (pp 24–33). Valencia, Spain.

- Scholman, M.C.J. and Demberg, V. (2017). Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue & Discourse, 8*(2), 56–83.

- Scholman, M.C.J., Rohde, H. and Demberg, V. (2017). "On the one hand" as a cue to anticipate upcoming discourse structure. *Journal of Memory and Language, 97*, 47–60.

- Scholman, M.C.J., Demberg, V. and Sanders, T.J.M. (submitted). When *for example* and *specifically* can be interpreted as *because*: Individual differences in coherence relation interpretation biases. *Submitted to Discourse Processes.*

Additionally, work on topics relevant to this dissertation have been published in the following articles:

- Evers-Vermeul, J., Hoek, J. and Scholman, M.C.J. (2017). On temporality in discourse annotation: Theoretical and practical considerations. *Dialogue & Discourse, 8*(2), 1–20.
- Hoek, J. and Scholman, M.C.J. (2017). Evaluating discourse annotation: Some recent insights and new approaches. *Proceedings of the 13th Joint ACL - ISO Workshop on Interoperable Semantic Annotation* (pp 1–13). Montpellier, France.
- Scholman, M.C.J., Evers-Vermeul, J., and Sanders, T.J.M. (2016). A step-wise approach to discourse annotation: Towards a reliable categorisation of coherence relations. *Dialogue & Discourse, 7*(2), 1–28.

# Chapter 2

# Background

This chapter delineates some of the basic concepts associated with the work reported in this thesis: the cognitive status of coherence relations, the segments between which a relation holds, which types of relations are distinguished, and how relations are constructed. The chapter also gives an overview of methods investigating discourse coherence, such as annotation studies, comprehension and processing experiments, and crowdsourcing studies.

These basics about coherence relations and their construction, as well as relevant methodologies for investigating coherence are relevant for the remaining chapters in this dissertation. Individual chapters also provide more detailed background information regarding some of these topics.

## 2.1 Coherence relations as part of the mental representation

When comprehenders understand a discourse, they understand more than what is explicitly stated in the text: they engage in inferencing processes to create connections between the individual elements and assign meaning to the text. In other words, comprehenders construct a *coherent mental representation* of the text. A mental representation is created by constructing *coherence relations* (cf. Hobbs, 1979; Mann & Thompson, 1988; Sanders, Spooren & Noordman, 1992), also referred to as discourse relations (cf. Asher, 1993; Lascarides, Asher & Oberlander, 1992; Prasad et al., 2007).

The meaning of coherence relations "cannot be described in terms of the meaning of the segments in isolation" (Sanders et al., 1992, p. 2). In other words, the meaning is more than the sum of its parts. This meaning is not an inherent property of the text; rather, it is assigned by the comprehenders.[1] The relations that comprehenders

[1] The writer may have intended for readers to construct a specific relation, but it is up to readers to infer the relation.

construct are a product of their psychological representations and processes. Coherence relations are therefore cognitive entities that are constructed in the minds of readers. To illustrate this, consider Passage (4), taken from a 2003 New York Times article discussing the success of the iPod.[2]

(4) [...] I made a jokey observation that before long somebody would probably start making white headphones so that people carrying knockoffs and tape players could fool the world into thinking they had trendy iPods. Jobs shook his head. "But then you meet the girl, and she says, 'Let me see what's on your iPod.' **You pull out a tape player, and she walks away.**"

The two clauses in bold are linked by *and*, which indicates a logical conjunction between the two segments (that is, two events occur without an explicit causal connection between them). However, to fully understand this text, comprehenders assign meaning to it that is not explicitly stated; that is, when reading the text, readers likely infer additional content, and then infer the connection between the clauses: when the boy would pull out a tape player, the girl would be disappointed that it was not an iPod, so she would walk away. In other words, the girl would walk away *because* the tape player is not a real iPod. This causal connection between the two clauses is created by readers during the interpretation process and is therefore part of the mental representation of the text; not of the text itself.

## 2.2 The segments of a coherence relation

Coherence relations hold between two text spans, which are referred to as segments or arguments. The segments of a relation are the "idea units" between which a relationship holds. They can vary in size and grammatical structure. Although a significant amount of research has aimed to characterise the set of relations that can hold between segments, less attention has been given to the characterisation of segments (but see Hoek, Evers-Vermeul & Sanders, 2017a). Proposals prescribe different segmentation rules, which has resulted in disagreement regarding what text spans constitute relational segments.

Often, the smallest unit that can function as a relational segment is taken to be the grammatical clause, which is a unit headed by a verb (Mann & Thompson, 1988; Prasad et al., 2007; Wolf & Gibson, 2005). Full sentences are generally also considered segments, but theories differ in where they draw the line regarding longer units. Some proposals focus on local relations, thereby only considering relations that hold between adjacent clauses and sentences within paragraphs.[3] Other theories also consider global

[2] Source: Rob Walker (2003, November 30th). The guts of a new machine. *New York Times.*

[3] Note that this might be due to theoretical reasons (i.e., it is assumed to relations between larger spans are not coherence relations but rather discourse structural relations) or practical reasons (i.e., agreement on relations with larger spans might be more difficult to reach).

relations; that is, relations between multiple sentences and paragraphs. Such relations make up the general text structure. In this thesis, the segments of relations under investigation range in size from clauses to paragraphs. The size of segments is mainly relevant in the first few chapters, where different proposals are compared.

Another area of discussion regarding segments is the structure that they adhere to; that is, where relations are needed in a coherent discourse (Knott, 1996). Again, theories vary in how they define segment structure. Some proposals prescribe that a discourse structure must be fully connected so that there are no segments of a text that are not part of a coherence relation with another segment; others do not have this expectation. Moreover, there are theories that allow for multiple relations to hold between two segments, but many proposals do not account for multiple relations. Still other proposals assume that one relation holds between every segment in a text, and that these segments can then be connected recursively until one high-level relation between paragraphs is identified. Such proposals do not allow for discontinuous segments. Chapter 3 elaborates on the structure of the segments prescribed by different proposals. This also becomes relevant in the final chapters, which examine whether comprehenders are sensitive to discourse structure when processing coherence.

## 2.3 Types of coherence relations

In the field of discourse coherence, it is generally assumed that the relations between segments can be characterised into a fixed, limited number of types. Proposals that detail which specific types are assumed to hold are referred to as "frameworks." At a coarse-grained level, frameworks often distinguish four main types: causal, additive, temporal and adversative relations. Causal relations are characterized by an implication relation between the segments: one segment presents the cause, and the other presents a consequence. Additive relations are logical conjunctions. Temporal relations are those where real events described in the segments are ordered in time (but do not have a causal relationship between them). Finally, adversative relations are characterized by the segments being in contrast with each other. The following examples – taken from a book describing the daily rituals of philosophers, writers, composers and artists – illustrate these relation types.[4]

(5) [F. Scott Fitzgerald wrote in gin-fuelled bursts]$_{\mathrm{S1}}$ – [he believed alcohol was essential to his creative process.]$_{\mathrm{S2}}$

(6) [Benjamin Franklin took daily naked air baths]$_{\mathrm{S1}}$ and [Toulouse-Lautrec painted in brothels.]$_{\mathrm{S2}}$

(7) Literary legend has it that [Edith Sitwell used to lie in an open coffin for a while]$_{\mathrm{S1}}$ before [she began her day's work.]$_{\mathrm{S2}}$

[4] Source: Mason Currey, Daily Rituals. Examples taken from the cover and page 148.

(8) [Freud worked 16 hours a day,]$_{S1}$ but [Gertrude Stein could never write for more than 30 minutes.]$_{S2}$

Example (5) presents a causal relation: the reason that Fitzgerald wrote in gin-fuelled bursts is that he believed it facilitated his creative process. Example (6) presents an additive relation, describing the daily ritual of two individuals. Example (7) presents a temporal relationship: Sitwell would presumably first lie in an open coffin, and then begin her work. Finally, Example (8) juxtaposes the writing habits of Freud and Stein, and the relation is therefore an adversative one.

Most researchers in the field of discourse coherence agree on the existence of these coarse-grained types of relations. However, researchers often disagree on a set of finer-grained relations and, as a result, there is no uniform set of relations that the community has agreed on (Hovy & Maier, 1995). Some proposals present sets of more than 70 relations (e.g., RST-DT, proposed by Carlson et al. 2003), others of only two relations (Grosz & Sidner, 1986). The proposals do not only differ in granularity, but also in the labels that they use: different labels are sometimes used for the same conceptual relations, and the same labels are sometimes used for different relation sense definitions. It is therefore not clear which and how many classes of relations and specific labels are needed to adequately describe and distinguish coherence relation types.

This thesis aims to investigate proposed distinctions between coherence relations. Specifically, Chapter 3 reviews some of the most well-known discourse frameworks, and discusses proposals to unify the relational labels of these frameworks. Chapter 4 then looks at how several frameworks relate to each other by considering data annotated using labels from multiple frameworks. These chapters reveal that there are significant differences between the inventories that are proposed by frameworks. In Chapter 5, we take a step back and evaluate how distinctions of coherence relations in these proposals can be justified. We argue that, ideally, the distinctions that discourse coherence experts have made in their inventories are justified by providing evidence from various sources (e.g., the linguistic system and experimental studies) using different methodologies, participants and materials. In Section 2.5, some of the main methodologies in discourse studies will be reviewed, to provide the background for the methodologies used in the current dissertation.

**Multi-level thesis** To complicate matters further, coherence relations can be described on different levels, corresponding to the primary function they perform in texts. Semantic information such as causality and temporal sequentiality is expressed by *ideational* relations, whereas the author's communicative goals such as to convince and explain are expressed by *intentional* relations. The ideational and intentional levels are also referred to as informational vs. intentional (Moore & Pollack, 1992), subject matter vs. presentational (Mann & Thompson, 1988), propositional vs. illocutionary (Sanders & Spooren, 1999), and ideational vs. interpersonal (Hovy & Maier, 1995). This debate will be referred to as the *multi-level thesis* (the term is proposed by Sanders & Spooren, 1999).

Both the ideational and the intentional function have their own corresponding relational labels: relations such as CAUSE or RESTATEMENT belong to the ideational group of relations, JUSTIFICATION or MOTIVATION belong to the intentional group (Hovy & Maier, 1995). Many relations that can be identified in a text can be assigned multiple labels, corresponding to both functions (Moore & Pollack, 1992). For example, instantiation (where one segment gives an example of something mentioned in the other segment) is a common way of providing evidence to support a claim (Carston, 1993). The INSTANTIATION relation is ideational, whereas the EVIDENCE relation is intentional.

Even though researchers often agree on the existence of these two functions, they disagree on how the functions relate to each other. Moore & Pollack (1992) assume that, in discourse interpretation, the recognition of one function follows from the recognition of the other function. Whether the ideational or argumentative function is recognised first depends on the relation, the context, and readers' knowledge. Mann & Thompson (1988) and Hovy (1995) assume that, for every relation, one function will be prevalent or primary given the relational context, and so one label will be paramount for that relation. Frameworks that propose an inventory of coherence relation labels deal with these functions in different ways.

The discussion of different levels of relations is elaborated in Chapters 7 and 8, which investigate whether readers are able to infer relations on both of these levels, and whether they display consistent biases in which type of relation they infer. In the next section, we turn to a discussion of the sources of information that comprehenders use to construct coherence relations.

## 2.4 Inferring coherence relations

Comprehenders can make use of different sources of information to infer coherence relations. The most well-studied type of cues that direct the interpretation process are connectives (such as *because* and *however*) and cue phrases (such as *for example* and *as a result*). Such cues provide comprehenders with "processing instructions" on how to connect incoming text inputs to previously read segments (Britton, 1994; Canestrelli, Mak & Sanders, 2013; Gernsbacher, 1997). Relations that are signalled by a connective or cue phrase are referred to as *explicit relations*.

Many relations are actually left *implicit*; that is, they are not marked by a connective or cue phrase (between 20-50% of relations, see Das & Taboada, 2018; Prasad et al., 2007). For such relations, comprehenders can exploit regularities of signals that tend to elicit expectations as to what relation will likely hold. Such signals include: lexico-semantic word pairs such as *good* – *bad* or *many* – *one of these* (Park & Cardie, 2012; Pitler, Louis & Nenkova, 2009), implicit causality verbs such as *ad-*

*mire* or *praise* (Ehrlich, 1980; Kehler, Kertz, Rohde & Elman, 2008; Koornneef & Van Berkum, 2006), transfer-of-possession verbs such as *give* or *bring* (Elman, Kehler & Rohde, 2006; Stevenson, Knott, Oberlander & McDonald, 2000), negation words such as *no* or *not* (Asr & Demberg, 2015; Webber, 2013), and tense (Das & Taboada, 2018). The most comprehensive collection of signals is the RST-DT Signalling Corpus (Das & Taboada, 2018), in which coherence relations have been annotated for all possible signals: reference, semantic, lexical, syntactic and graphical features.

Comprehenders can also make use of non-textual sources of information, such as background knowledge and world knowledge. These types of knowledge support the process of interpreting and representing the text (van den Broek, 2010). The identification of a relation often interacts with background knowledge: if comprehenders have knowledge about the topics of the sentences and the relations among these concepts, background knowledge will be activated and the inference will be checked against it (Noordman & Vonk, 2015). Consider Example (9), where S1 stands for "Segment 1" and S2 stands for "Segment 2."

(9) [I had a dial-up connection and a 24k modem.]$_{\mathrm{S1}}$ [It would take a day to download an album.]$_{\mathrm{S2}}$ *Source: Trevor Noah, Born a crime, p. 187*

The author of this example meant to express that it used to take a long time for him to download an album, and that the dial-up connection and the modem are the reasons for this. Comprehenders can arrive at this interpretation using their background and world knowledge: preferably, one would know that dial-up connections and 24k modems are not the state of the art when it comes to 21$^{\mathrm{st}}$ century computer network devices. Alternatively, one would have knowledge about how long it generally takes to download an album nowadays. Using this knowledge, it can be inferred that a day is a long time for downloading an album. Someone who does not have this knowledge might not be able to draw the inferences necessary to construct the intended causal relationship for this example.

Now consider Example (10).

(10) [I had a craving for traditional "Dibbelabbes".]$_{\mathrm{S1}}$ [It would take a day to travel from Utrecht to Saarbrücken.]$_{\mathrm{S2}}$

Example (10) is similar in structure to Example (9), but readers would likely not construct a causal relationship between these sentences. Readers with knowledge of *Saarlandisch* food would know that Dibbelabbes is a snack typical of the Saarland region. Even without this knowledge, readers might be able to infer that Dibbelabbes is available in Saarbrücken but not in Utrecht. They can then construct an adversative relation between these segments (*I had a craving, but it would take a day to satisfy it*) or an elaborative relation (*I had a craving, and it would take a day to satisfy it*). Hence, comprehenders are likely to construct different relations depending on the inferences they can draw and the background knowledge that they can depend on.

## 2.5 Methodology in studies of discourse coherence

The work in this dissertation is comprised of different methodologies – such as annotation studies, interpretation studies, story continuation studies, and eye-tracking studies – and this work is often placed in the context of existing work using yet other types of methodologies. In order to provide the necessary background, this section gives a brief overview of methodologies that are common in discourse studies investigating coherence relational categories.

Traditionally, researchers in the field of discourse investigate three aspects of language: acquisition, production, and comprehension (also see Sanders & Evers-Vermeul, to appear, for an overview on converging evidence). These three aspects are investigated using corpus-based data, as well as off-line comprehension and online processing experiments. We will briefly discuss each one in turn.

**Corpus-based data** In order to study the distribution and linguistic realization of coherence relations, researchers use discourse-annotated corpora, which are large-scale collections of texts containing coherence relation labels and other relevant information on how the sentences in the texts relate to each other. Corpora can contain texts from different genres (e.g., newspaper, novels, broadcast interviews) and modalities (e.g., spoken, written, chat). Researchers can use the annotated relations to test hypotheses based on distributional data, and to develop new hypotheses.

The most commonly used and comprehensive discourse-annotated corpora are manually annotated, since automatic parsers are not able to reliably classify implicit coherence relations as of yet (Park & Cardie, 2012; Pitler, Louis & Nenkova, 2009; Rutherford & Xue, 2014, 2015). In traditional discourse annotation, the standard practice is to employ two trained, expert annotators to code data (see also Spooren & Degand, 2010). Several issues can be identified with this coding procedure. First, manual annotation by experts is time-consuming and costly due to annotator training and the annotation itself (as discussed in, e.g., De Kuthy, Ziai & Meurers, 2016; Scholman, Evers-Vermeul & Sanders, 2016). Second, the procedure raises questions regarding the reliability and validity of the data (see also Hoek & Scholman, 2017; Scholman et al., 2016). Trained, expert annotators may agree not only because they carefully followed instructions, but also because they share implicit knowledge and know the purpose of the research well (Artstein & Poesio, 2008; Riezler, 2014). This type of agreement is not entirely valid. Krippendorff (2004) therefore notes that the more annotators participate in the process and the less expert they are, the more likely it is that the data is reliable (see also Riezler, 2014). Annotation of coherence relations by experts should therefore be complemented by external grounding.

**Experimental data** There are various methods of experimentally investigating discourse. Often, a distinction is made between off-line and online studies. Off-line measures provide insight into the representation of the discourse after the comprehension process is completed. Methodologies that tap into these representations include judgment tasks (participants are asked to provide acceptability or plausibility ratings of the discourse), recall (participants are asked to recall aspects of the discourse after having processed it), and comprehension questions (participants are probed to answer questions in order to test their interpretations). Another off-line measure that is used in several chapters of this dissertation is elicited production, specifically story continuation experiments. In such experiments, participants are presented with prompts and asked to continue the prompt in a natural manner. This type of study is useful to test whether readers are more likely to infer a specific type of relation depending on different contexts.

Although off-line measures provide valuable insight into readers' mental representation of the discourse, these measures do not capture the processes and representations that are constructed "online" during comprehension (Graesser, Millis & Zwaan, 1997). Different paradigms can be used to capture the on-line comprehension processes. Self-paced reading and eyetracking-while-reading are commonly used methods that allow the researcher to measure reading times as participants process a text. The reading times and eye movement data reflect moment-to-moment cognitive processes that occur during text processing (Rayner, 1998). Other online measures have a more neurological basis, such as electroencephalography (EEG) methods like event-related potentials experiments (ERP). During such experiments, participants are presented with auditory or visual texts while their brain responses (electrical activity at the scalp) to such stimuli are recorded. These methods provide a clearer view on what processes occur during reading; for example, they can reflect when participants update their expectations or are surprised by a semantic anomaly in the text.

**Crowdsourcing** Traditionally, the annotation of relations and collection of experimental data are done in laboratories, often using paper and pencil for off-line tasks and more advanced methods for online tasks. More recently, however, technological developments have made it possible to develop new ways of conducting experiments. Specifically, crowdsourcing platforms have become increasingly popular in the field of linguistics for generating data – whether annotations or experimental responses. These platforms function as online labor markets where participants perform small tasks in exchange for ethical rewards (i.e., participants are commonly rewarded at least $6.50 per hour) (Snow, O'Connor, Jurafsky & Ng, 2008). Crowdsourcing platforms provide a crucial method for collecting the data reported on in this dissertation. This section will therefore present a more detailed introduction to and discussion of crowdsourcing.

Among the most popular platforms are Amazon's mechanical Turk (which mainly consists of a US-based participant population) and Prolific (whose participant population has a more diverse linguistic background). Participants can sign up on these websites and enter demographic information regarding their first language, age, educational level, etc. Researchers can then target specific groups of participants easily and quickly.

The advantages of using crowdsourced participants are clear: it is easier to recruit participants in general and from different backgrounds in particular (compared to lab-based experiments, which are often conducted using college students as subjects), data can be collected faster and at a lower cost. Additionally, crowdsourcing is revolutionizing the way we do research by opening up new possibilities for investigating linguistic phenomena, as well as interpreting, representing, and analyzing data. It allows researchers to study the distribution of responses over many participants, rather than specific data points from a smaller sample size (see von der Malsburg, Poppels & Levy, 2018; Munro et al., 2010).

However, crowdsourcing discourse annotations, interpretations, and comprehension data also brings several possible pitfalls. In the lab we can meet our participants face-to-face and monitor them while they complete the study. Over the Internet, we cannot be so sure that our participants are who they say they are and that they are completing the study properly or are even paying attention. So, first, researchers need to do more quality controls for crowdsourced data to ensure that the data is not "noisy". Noisy data occurs when unmotivated participants try to maximize their pay by supplying quick answers without performing the task seriously. In order to mitigate the effect of such *spammers*, researchers typically collect multiple annotations so that they can later use de-noising methods to filter out the noisy data. Several de-noising methods can be used, ranging from simple majority voting to more advanced item-response models (see Hovy, Berg-Kirkpatrick, Vaswani & Hovy, 2013; Passonneau & Carpenter, 2014). By including enough participants and using de-noising methods, researchers can obtain data of high quality (comparable to data obtained in laboratories) (see, e.g., Munro et al., 2010; Snow et al., 2008).

Second, tasks need to be designed in a way that ensures participants cannot cheat, for example by creating speeded tests. Consider the Author Recognition Test (ART, Stanovich & West, 1989), which is a simple test that is commonly used in lab-based settings as a measure of linguistic experience. Participants are presented with a list of names and asked to indicate which names they recognize as author names. In a crowdsourcing setting, participants might be enticed to look up certain names so that they will perform well. In order to ensure that participants cannot do this, the copying of the author names on the webpage can be blocked and time-out features can be added to such studies, in which case participants only have a few seconds per item to provide an answer.

Another possible pitfall concerns the use of crowdsourcing methodologies for discourse annotation studies. Non-expert workers have been used in a variety of linguistic annotation tasks, such as affect recognition, word similarity, audio segmentation, and focus annotation (De Kuthy et al., 2016; Munro et al., 2010; Snow et al., 2008), but crowdsourcing has rarely been used to obtain discourse-annotated data. This could be due to the nature of crowdsourcing: typically, crowdsourced tasks are small and intuitive tasks. Under these conditions, crowdsourced workers – unlike expert annotators or in-lab naïve annotators – cannot be asked to work according to a specific discourse theory because this would require them to receive training first. Since it is not straightforward to train crowdsourced workers to use a specific approach, it might seem as if crowdsourcing is not suitable for obtaining coherence relation annotations. However, for the research in the current thesis, we designed a task in such a way that somebody who is not an expert in linguistics is able to provide annotations without receiving much training. This is particularly valuable because it allows researchers to obtain many different interpretations of a single relation – thereby revealing all the meanings of that relation, rather than only revealing the dominant one, which is currently represented in corpora.

Collecting off-line experimental data using crowdsourced participants is more straightforward than collecting discourse annotations. Off-line tasks can easily be designed as short and simple tasks, making them very suitable for crowdsourcing. Online tasks are less suitable for crowdsourcing, however, because many methods require equipment such as eye-tracking cameras and EEG electrodes. In Chapters 9 and 10, we test several hypotheses by complementing online eye-tracking experiments conducted in laboratory settings with off-line crowdsourced experiments, and examining these results side by side.

## 2.6 Summary and conclusion

The first part of this chapter further defined the notion of coherence relations, discussing in particular which types of coherence relations are taken to exist and how relations can be inferred. Additionally, this chapter posited that coherence relations are cognitive entities that are constructed in the minds of readers.

The second part of this chapter introduced various methodologies that are common in discourse studies and relevant to the work in this dissertation. Corpus-based data is most relevant to Chapters 3 and 4, whereas experimental data is mainly relevant to Chapters 6-10. Chapter 5 discusses why these different methodologies are all crucial for justifying relational distinctions, and makes a plea for combining various methodologies to provide converging evidence for theories. Finally, the current chapter provided a more detailed explanation of crowdsourcing, including its advantages and disadvantages. This discussion will become relevant starting from Chapter 6.

# Chapter 3

# Unifying coherence relation frameworks[1]

Proposals of coherence relational inventories differ greatly in the number and type of coherence relations that are distinguished, with different frameworks varying from sets of approximately 20 relations (such as the original RST developed by Mann & Thompson 1988), to others of only two relations (Grosz & Sidner, 1986). Consequently, the corresponding corpora are not interoperable: researchers working in one framework cannot easily make use of data annotated according to a different framework. In this chapter, we discuss several proposals for unifying the different frameworks (either by creating a new intermediate framework or by proposing an intermediate representation for relational types). We then introduce the Unifying Dimensions approach (UniDim: Sanders, Demberg, Hoek, Scholman, Torabi Asr, Zufferey & Evers-Vermeul, 2018), which forms the basis for the research described in this dissertation. The UniDim proposal allows researchers to make use of data from different frameworks by decomposing the labels into values for several dimensions. These dimensions are taken from the Cognitive approach to Coherence Relations (Sanders et al., 1992), and are combined with more fine-grained additional features from the frameworks themselves. This yields a posited set of dimensions that can successfully map several frameworks.

## 3.1 Introduction

A substantial number of efforts have been made to develop an inventory of coherence relations, which has led to the existence of a variety of different coherence frameworks

[1] The Unifying Dimensions mapping that is discussed in this chapter was proposed in Sanders, T.J.M., Demberg, V., Hoek, J., Scholman, M.C.J., Asr, F.T., Zufferey, S., and Evers-Vermeul, J. (2018). Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory, ahead of print*, 1–71. The discussion of the mapping in this chapter is based on, and in some parts identical with, this article.

and corresponding corpora. These frameworks differ greatly in the type and number of coherence relations that are distinguished, and in the definition of labels (i.e., different labels are used for the same conceptual relations, and the same labels are used for different relation sense definitions).

The lack of interoperability between frameworks makes it difficult to draw comparisons across the corresponding corpora. This is an issue, because interoperability would allow researchers working with a particular framework to use data from other frameworks for investigating certain coherence phenomena (for example, to investigate which cues besides connectives typically signal relatively uncommon relations like Chosen alternative). Furthermore, interoperability of corpora would be useful for researchers working on automatic coherence relation classification. Many natural language processing tasks, such as information retrieval and question-answering systems (Jansen, Surdeanu & Clark, 2014; Sharp, Jansen, Surdeanu & Clark, 2015), text summarization systems (Louis, Joshi & Nenkova, 2010; Marcu, 2000), and machine translation systems (Koehn, 2009; Meyer & Popescu-Belis, 2012) would improve from increased performance in automated relation classification. Current state-of-the-art systems (see  Xue et al., 2015, for an overview) make use of manually annotated corpora for training, but the lack of training data poses problems for the accuracy of these systems (see also Shi & Demberg, 2017). The performance of these tools would likely increase if more training data could be used. Interoperable corpora would allow researchers to train their models on multiple corpora, rather than just those corresponding to a specific framework.

To address the issue of interoperability, several efforts have been made to unify different frameworks by describing how the relational labels relate to each other. This chapter discusses these efforts and introduces a new proposal. First, we provide a general description of some of the major frameworks.

## 3.2 An overview of discourse annotation frameworks relevant to this dissertation

This section provides information on several discourse frameworks, with a focus on those that are most relevant to the work described in this thesis; however, many more relevant discourse annotation frameworks exist (e.g., SDRT, Asher & Lascarides 2003; GraphBank, Wolf & Gibson 2005; among many others). The frameworks that play a central role in this thesis are those used to annotate the Penn Discourse Treebank (PDTB) and the Rhetorical Structure Theory Discourse Treebank (RST-DT), as well as the Cognitive approach to Coherence Relations (CCR). For each framework, some of the basic premises and the (structure of the) relational inventory are described.

### 3.2.1 PDTB

The PDTB 2.0 corpus (Prasad et al., 2008) is the largest manually annotated discourse relation corpus available at the moment. The framework that was used to annotate the corpus has also been used to create new corpora in other languages and genres, resulting in different styles of PDTB annotation (but they can be considered to be interoperable; cf. Prasad et al., 2014). Here, we use the term PDTB to refer to the framework that corresponds to the original PDTB-style annotation. The PDTB Research Group plans to release a new version of the corpus (PDTB 3.0) (see also Webber, Prasad, Lee & Joshi, 2016). The differences between the versions are described where relevant.

#### Basic premises

PDTB is characterized by two basic premises. First, the PDTB follows a lexically-grounded approach to discourse relation representation, meaning that they focus on annotating lexical items that can signal discourse relations. The framework is also characterized by its theory-neutral approach to annotation, thereby making no commitments to what kinds of high-level structures may be created from the low-level annotations of relations and their arguments (Prasad et al., 2008).

Discourse relations are taken to hold between two arguments, referred to as Arg1 and Arg2 according to syntactic conventions, and are triggered by connectives or by adjacency between clauses (Prasad et al., 2017). The PDTB distinguishes between explicit and implicit relations. Explicit relations are marked by an explicit connective. The arguments of explicit relations are unconstrained in terms of their location. The argument that is syntactically bound to the connective is always labeled as Arg2; the other argument is Arg1. Annotators are instructed to annotate the relation label that corresponds to the connective; in case a relation is doubly-marked (e.g., *but also*), both connectives are annotated separately.

For implicit relations, annotators are instructed to insert a suitable connective and then annotate the corresponding relation label. Implicit relations are only annotated for adjacent sentences within paragraphs (see also Prasad et al., 2017), as well as between complete clauses delimited by a semi-colon (";") or colon (":"). Arg1 and Arg2 reflect the linear order of the arguments, with Arg1 always being the first sentence.

Three distinct labels were used for cases where an implicit connective could not be provided. AltLex (alternatively lexicalized) applies to relations for which insertion of a connective leads to a perception of relation redundancy, indicating the presence of an alternative lexico-syntactic marking of the relation (Prasad et al., 2017). AltLex relations hold only between two adjacent arguments. EntRel is used for entity-based relations, which hold between two adjacent arguments for which no discourse relation can be inferred. NoRel is annotated when no discourse or entity relation holds between two adjacent sentences.

PDTB 2.0 allows annotation of multiple labels for a single connective, but this rarely occurred (5.4% of explicit relations, Prasad et al. 2014). However, PDTB 3.0 will have more multiple-labeled instances (see Webber et al., 2016).

The annotations in the PDTB adhere to the *minimality principle*, which requires an argument to contain only the minimal amount of information needed to complete the interpretation of the relation (Prasad et al., 2008). Any other span that is considered relevant but not necessary to the relation is annotated as supplementary information.

### The (structure of the) relational inventory

PDTB's relational inventory consists of 43 relation labels that are presented as a hierarchical classification scheme with three levels: *class*, *type*, and *subtype* (the PDTB 2.0 and 3.0 hierarchies are added as Appendix A).[2] The top level, *class*, distinguishes four major semantic classes: TEMPORAL, CONTINGENCY, COMPARISON, and EXPANSION. These classes are further refined in types. For example, the class CONTINGENCY contains relation labels for different CAUSE and CONDITION relation types, and COMPARISON is specified in different CONTRAST and CONCESSION labels. The third level, subtype, specifies the semantic contribution of each argument. For example, CONCESSION has two subtypes: EXPECTATION (Arg1 denies the expectation created by Arg2) and CONTRA-EXPECTATION (Arg2 denies the expectation created by Arg1).

The hierarchical structure of the inventory benefits the inter-annotator agreement (Prasad et al., 2008): when annotators disagreed on the label for a certain relation (e.g., OPPOSITION VS. JUXTAPOSITION), the label could be adjudicated to the higher common label (in this example, CONTRAST). Similarly, when annotators were not sure of which fine-grained sense to choose, they could choose a higher relation label.

The PDTB 2.0 annotation scheme and relational inventory is described in detail in the manual (Prasad et al., 2007). The PDTB 3.0 relational inventory differs slightly from the PDTB 2.0's inventory. A preliminary version of PDTB 3.0's scheme is described in Webber et al. (2016). The top-level classes remain the same, but several adjustments have been made at the type and subtype levels. Some existing classes were refined; for example, the class CONTINGENCY now also contains type labels for CAUSE-BELIEF relations. Some subtype labels were removed, such as CONTRAST'S subtypes OPPOSITION and JUXTAPOSITION. Moreover, new labels were added, such as the type label MANNER. The current PDTB 2.0 corpus is being reannotated with PDTB 3.0 labels; this updated corpus has not been released as of yet.

[2] The respective levels of labels are indicated using capitalization: class labels are written in all caps, type is indicated with the first letter capitalized, and subtype labels are written in small caps.

### 3.2.2 RST-DT

The framework that is used to annotate the Rhetorical Structure Theory Discourse Treebank (RST-DT, Carlson & Marcu, 2001) is based on the Rhetorical Structure Theory (RST), originally proposed by Mann & Thompson (1988). There are many implementations of RST-style annotation, but in this dissertation, we will focus on the style used to create the RST-DT.

**Basic premises**

RST-DT is a descriptive theory of the relations that hold between text segments, originally developed to guide computational text generation (Taboada & Mann, 2006). Relations in RST-DT can be of semantic, intentional, or textual nature (Carlson, Marcu & Okurowski, 2003). The intentional nature of relations corresponds to the writer's intentions: annotations relate to the writer's goal or intended effect of each segment of a text with respect to the neighboring segments. RST-DT allows for the annotation of only one label for every instance.

Relations are defined to hold between two (or more) non-overlapping text spans. An essential characteristic of RST-style annotation is the assignment of nuclearity: texts spans are characterized as nuclei or satellites (every relation has to consist of at least one nucleus). The nucleus is the central part of a text (with respect to its intentional discourse structure), while the satellite is supportive of the nucleus. Some relations have symmetrically important arguments by definition. These relations consist of two nuclei rather than a nucleus and a satellite, and are referred to as *multinuclear* relations. The writer's intentions are important when assigning the nuclearity (i.e., what does the writer want to achieve?). Determining nuclearity can therefore rarely be done without taking the context of the relation into consideration. Nuclearity assignment is determined simultaneously with the assignment of a discourse relation (Carlson et al., 2003).

A fundamental constraint in the RST framework is that the discourse structure of a text has to be represented as a hierarchical tree structure. The leaves of the tree correspond to the *elementary discourse units* (EDUs) – the minimal segments in a text. EDUs generally consist of clauses, but attributions, relative clauses, nominal postmodifiers, and phrases that begin with a strong discourse marker are also considered EDUs in RST-DT (see Carlson & Marcu, 2001, p.3). Determining the EDUs of a text comprises the first step in RST-style annotation. Next, nuclearity is assigned to the nodes and adjacent spans are linked together via rhetorical relations. These relations are linked recursively, thereby creating a hierarchical tree structure.

**The (structure of the) relational inventory**

The RST-DT framework consists of 78 relation labels (see Appendix B), which are partitioned into 16 classes that share some rhetorical meaning such as Cause or

ELABORATION. The inventory is data-driven, based on analysis of the RST-DT corpus, which consists of Wall Street Journal articles. Some of the relational classes contain relations that are not considered to be coherence relations in other approaches such as PDTB 2.0; examples include the ATTRIBUTION (which is not considered a coherence relation in PDTB, but rather annotated as additional information) and QUESTION-ANSWER relations. Another example is cases where coherence between discourse segments is not achieved through a specific semantic coherence relation but rather through cohesion (ENTREL in PDTB). These cases are annotated as ELABORATION relations in RST-DT.

Relational definitions are based on functional and semantic criteria, not on signals, because the creators argued that no unambiguous signal for any relation was found (Taboada & Mann, 2006); i.e., a connective such as *but* can mark different types of relations. RST-DT differs from PDTB in this respect by not following a lexically grounded approach.

### 3.2.3 CCR

The Cognitive approach to Coherence Relations was proposed as a psychologically plausible theory of discourse structure and coherence relations (Sanders et al., 1992). Only later was it applied in discourse annotation efforts (see Sanders, Vis & Broeder, 2012; Scholman, Evers-Vermeul & Sanders, 2016; also see Hoek, 2018, for an overview of CCR's applications to discourse annotation). To date, no English CCR corpus exists that is comparable in size to PDTB or RST-DT.

#### Basic premises

CCR's relational distinctions are based on two requirements: descriptive adequacy and psychological plausibility. Descriptive adequacy requires the theory to describe the structure of all kinds of natural texts. Psychological plausibility requires the theory to generate plausible hypotheses on the role of discourse structure in the construction of the cognitive representation. Such a claim leads to the prediction that coherence relations and their linguistic marking affect the cognitive representation of a discourse. Consequently, the distinctions that are made in a relational inventory must be supported by evidence from processing, comprehension and/or acquisition studies in order to be considered cognitively plausible. The theory strives to function as an economic theory that generates a limited set of classes of coherence relations, which can be extended by using segment-specific features (Sanders et al., 1992).

CCR considers the segments of a relation to be minimally clauses. Following the surface order in which they appear, CCR refers to the first segment as S1 and to the second segment as S2. It is assumed that each segment expresses an underlying proposition. The propositions can be events, states, speech acts, claims, opinions and judgments. The underlying propositions of segments S1 and S2 are referred to as 'P' and ‘Q’ in CCR (as in propositional logic), where P represents the antecedent of a relation and Q represents the consequent of a relation. The coherence relation is then defined by the logical relation between P and Q, and by the way in which S1 and S2 map onto these propositions. It is therefore assumed that S1 and S2 directly or indirectly express the underlying propositions P and Q. To illustrate this, consider the following examples:

(11) [Billy yelled at his nanny,]$_{\mathrm{P}}$ so [she gave him a time-out.]$_{\mathrm{Q}}$

(12) [Billy yelled at his nanny]$_{\mathrm{Q}}$ because [she gave him a time-out.]$_{\mathrm{P}}$

The segments of these examples are similar, but the logical relation between P and Q is reversed. In Example (11), S1 presents the cause (or antecedent) and S2 presents the consequence of the proposition in S1. P therefore precedes Q in this example; it is a cause-consequence relation. In Example (12), however, the consequent precedes the antecedent: S1 maps onto the consequence (Q), and S2 onto the cause (P). The example is therefore a consequence-cause relation.

### The structure of the relational inventory

Instead of using classes and end labels to distinguish types of relations, CCR decomposes relations into basic dimensions that hold for every relation. Originally, CCR distinguished four dimensions; more recently, a fifth dimension has been proposed (Evers-Vermeul, Hoek & Scholman, 2017). Considering all possible combinations of values for these five dimensions leads to a taxonomy of classes of relations.

The first dimension refers to the *polarity* of relations: it distinguishes between `positive` and `negative` relations. A relation is `positive` if the relation holds between P and Q. Typical markers for `positive` relations are *and* or *because.* A relation is `negative` if a negated version of P or Q functions in the relation. Typical connectives are *but* and *although.* Polarity distinguishes adversative relations such as CONTRAST and CONCESSION from all other relations.

The second dimension is the *basic operation.* The basic operation dimension has two values: `causal` and `additive`; based on the assumption that only these two kinds of relations can exist between P and Q (Sanders et al., 1992). A relation is `causal` if an implication relation (P → Q) can be deduced between the two segments. Typical connectives are *because* and *although.* A subtype of causal relations is `conditional`: when the event or situation in one of the segments has not been realized yet. A relation is `additive` if the segments cannot be connected by an implication relation, but rather in a logical conjunction (P & Q). Typical connectives are *and* and *then.*

The third dimension is the *source of coherence*, which distinguishes between `objective` and `subjective` relations.[3] A relation is `objective` if both segments describe situ-

[3] Originally, the terms `objective` and `subjective` were defined in the literature as `semantic` and `pragmatic`, respectively (Sanders et al., 1992, 1993; Sanders, 1997; also see Pander Maat & Sanders, 2000 for a discussion of this transition).

ations in the real world. The speaker merely reports these facts and is not actively involved in the construction of the relation. Relations are subjective if speakers or authors are actively engaged in the construction of these relations, for example because they are reasoning or concluding.

Several annotation studies have shown that the source of coherence can be difficult to determine (see, for example, Sanders et al., 1992; Sanders, 1997; Scholman et al., 2016; as well as the annotation effort presented in the next chapter). In order to facilitate the annotation of source of coherence, coders can make use of paraphrase tests, in which the segments of the relation are inserted in a paraphrase that makes explicit whether the relation is `objective` or `subjective`. For example, coders can paraphrase causal relations with a basic order as *the fact that P causes the speaker's claim/advice/conclusion that Q* (Sanders, 1997; Scholman et al., 2016; see also Hoek, 2018).

The fourth dimension is the *implication order of the segments* (also known as *implication order*) and refers to the order in which P and Q are expressed in the discourse. For example, if a cause (P) precedes the consequence (Q) in a causal relation, the relation is said to be of `basic` order; if the consequence precedes the cause, the relation has a `non-basic` order. `Additive` relations are by nature symmetrical, and therefore the order dimension does not apply to these relations.

The fifth dimension is *temporality*, which defines whether two segments are ordered in time. When S1 and S2 display a temporal sequence of temporal overlap, they are `temporally ordered`. If temporal order is not relevant to the relation at hand (as in purely `additive` relations), the relation is `non-temporal`.

## 3.3 Creating a unified framework

The existence of many different frameworks that make use of various relational categorizations creates complications for the interoperability of these frameworks: it is difficult to compare the annotations done using different frameworks since it is not clear how relational labels map onto each other. Different labels are used for the same conceptual relations, and the same labels are used for different relation sense definitions. Benamara & Taboada (2015) list four specific ways in which relational categories in different frameworks tend to overlap or be related. First, *Specialization* occurs when a relational label in one approach can correspond to several labels in another approach. It's opposite, *Generalization*, occurs when several labels in one approach correspond to one label in another approach. *Omission* refers to a relational label that is distinguished in one approach, but not taken into account in another. Finally, *Definition* refers to relational labels that have similar names, but different definitions.

In order to improve the interoperability of various frameworks and their corresponding resources, the exact links between the labels and classes of different frameworks need to be established. Researchers have gone about this in various ways. In this section, we look at several proposals for relating frameworks to each other, before turning to a new proposal that is a central part of the work reported in this thesis: the Unifying Dimensions proposal.

### 3.3.1 Related work

Several researchers have set out to develop a single unified framework based on the classes from multiple frameworks. One of the first proposals was put forward by Hovy & Maier (1995). They note that many different proposals exist for discourse relation inventories, and aim to provide a set that unites the frameworks. In order to create a common set, they taxonomized the more than 400 relations that have been proposed in over 30 different frameworks into a hierarchy of roughly 70 discourse relations. Their taxonomy is two-dimensional: one dimension is constrained in the number of relations – the general relation types, which are higher in the hierarchy – and the other is unconstrained – the more specific labels, which are lower in the hierarchy. The top tier consists of three broad groups: *ideational* (also known as semantic), *interpersonal* (i.e., author- and/or addressee-related), and *textual* (i.e., presentational). Semantic information such as causality and temporality is expressed by ideational relations; interpersonal relations express the author's communicative goals such as motivating and describing; and textual relations are used to form the discourse into a coherent whole (Hovy & Maier, 1995). Although no specific one-to-one mapping between Hovy and Maier's taxonomy and the PDTB and RST-DT exist (the two approaches did not exist yet), it is likely that a mapping can be created since the taxonomy is based on general classes of relations found in over 30 different frameworks.

Bunt & Prasad (2016) also proposed a general set of relations: they developed an international standard (ISO standard) for coherence relation annotation, which consists of a set of 20 "core" relations that are commonly found in some form in existing approaches. They did not aim to provide a fixed and exhaustive set of coherence relations; rather, they aimed at providing an open, extensible set of relations. Bunt and Prasad (2016) propose that the ISO standard can be used for future annotation efforts, as well as for mapping between annotations using different frameworks. To this end, they provided mappings of these ISO relations to other frameworks, including PDTB and RST-DT. From these proposed correspondences to ISO standard candidate relations, we can infer how the PDTB 2.0 and RST-DT relations correspond to one another.

Another proposal for creating a new framework comes from Benamara & Taboada (2015). They developed a unified hierarchy of 26 types of discourse relations based on RST and SDRT, which can be used to annotate new corpora as well as map existing annotations. They aim at making their hierarchy stable enough for language variation and open to modification at the low level. The resulting taxonomy is intentionally and semantically driven. The three-tiered hierarchy has four top-level classes: Tem-

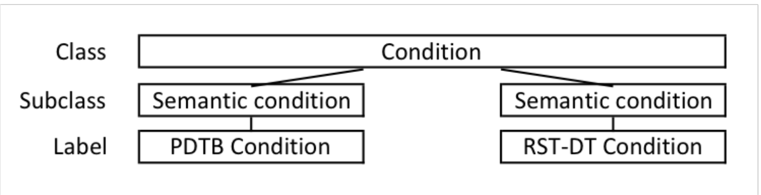


**Figure 3.1:** Illustration of OLiA's mapping approach.

poral, Structural, Thematic, and Causal-argumentative. They mapped their taxonomy to annotations in three corpora (RST-DT, RST-ST, and SDRT Annodis), but they did not have any data available that was annotated according to both frameworks. They were therefore not able to evaluate whether the actual annotations of the two frameworks were consistent with each other. Moreover, as of yet, there is no mapping between their taxonomy and PDTB's labels available.

Benamara & Taboada (2015)'s work highlights differences in granularity between frameworks by identifying certain labels that exist in one framework but do not have a corresponding label in the other framework (i.e., cases of *omission*). Such relations with no correspondence across taxonomies need more consideration when using an intermediate framework: if one of the frameworks does not include a relation label that is included in another framework, the data for those instances cannot be mapped. Ignoring them leads to an incomplete mapping, but including them would mean additional manual annotation of these instances. To illustrate this, consider PDTB's Exception relation, which is mapped to ISO's Exception relation. No direct equivalent for Exception exists in RST-DT, so no label is mapped to this particular ISO label. This means that there is no correspondence between PDTB's Exception and any label in RST-DT, but it can be assumed that in RST-DT, these relations might be given a more general label. Mapping using an intermediate framework therefore has as a downside that it cannot easily provide a one-to-many mapping in case there is no direct corresponding label in one of the frameworks.

Creating a new, intermediate framework is not the only way to unify existing frameworks. Another solution for how to improve interoperability is to create an intermediate *representation* for mapping between frameworks. Chiarcos (2014) was the first to attempt this. He mapped the PDTB and RST-DT relation labels onto each other as part of the Ontologies of Linguistic Annotation (OLiA). OLiA provides a terminology repository that can be used to facilitate the conceptual interoperability of annotations. This is done using an intermediate level of representation that mediates between several existing frameworks. The intermediate representation is formalized as *subClassOf* descriptions. This approach is illustrated in Figure 3.1, which shows the mapping of PDTB's Condition to RST-DT's Condition. In OLiA, this relation type is characterized as a subclass of *Semantic condition* relations, which is in turn a subclass of *Condition* relations. This can then be mapped onto RST-DT's class of *Condition*, which has the same superclasses. Chiarcos (2014) argues that ontologies are able to represent more fine-grained nuances of meaning, and to quantify the number of shared descriptions between annotations of different frameworks.

OLiA's *subClassOf* intermediate representation theoretically allows all relational labels to be mapped, even if there is no direct correspondence: one can simply map the labels to an overlapping superclass. However, the correctness of mapping to a higher overlapping superclass is not always guaranteed and can lead to very general mappings. To illustrate this, consider RST-DT's CIRCUMSTANCE label – used for relations where one segment provides the context for the other segment, and the situations in the segments are somewhat co-temporal. The superclass for this relation in OLiA's mapping is *Background*, but such a superclass does not exist in PDTB. The highest overlapping class (i.e., the OLiA class that applies to labels prescribed by both RST-DT and PDTB) is the supersuperclass *Expansion*. Mapping RST-DT's relational label CIRCUMSTANCE to the PDTB class of EXPANSION relations would be uninformative because this class is very general, covering a set of 10 different relation types. Moreover, mapping RST-DT CIRCUMSTANCE to the PDTB class EXPANSION ignores the temporal aspect of CIRCUMSTANCE relations: relation types belonging to PDTB EXPANSION are additive and not temporal. This could create difficulties for the mapping in practice if an instance annotated as RST-DT CIRCUMSTANCE is annotated as a temporal relation in PDTB – this would appear to be a "false" mapping, even though it could be valid.

### 3.3.2 Unifying Dimensions: a new proposal

In order to be able to represent all annotations that the different schemes have considered relevant for discourse relation annotation, we proposed a new method to map relation labels onto each other using an intermediate interlingua (Sanders, Demberg, Hoek, Scholman, Torabi Asr, Zufferey & Evers-Vermeul, 2018). We compared PDTB and RST-DT, as well as a third framework (SDRT), in terms of a limited set of dimensions, and showed how they map onto each other. For instance, all systems distinguish between adversative and non-adversative relations, known as `positive` and `negative` relations in UniDim. We then showed how this dimension allows for similar clusterings across frameworks. The dimensions that are part of the Unifying Dimensions proposal come from the Cognitive approach to Coherence Relations (CCR), while additional criteria were added to capture more fine-grained distinctions. All relation labels from the included frameworks were decomposed for their values on these dimensions and features, which revealed how the various proposals can be related to each other. This approach allowed us to "translate" labels from one framework to the terminology of another. The ultimate goal of this proposal was to make optimal use of existing corpora and facilitate discussion among researchers working with different frameworks.

### Dimensions included in UniDim

The starting point of UniDim was the dimensions included in CCR: polarity (`positive` vs. `negative`), basic operation (`causal` – including `conditional` – vs. `additive`), source of coherence (`objective` vs. `subjective`), order of the segments (`basic` vs. `non-basic`), and temporality (`synchronous`, `asynchronous` or `non-temporal`). These dimensions were applied to relation labels from PDTB, RST-DT, and SDRT. The translation of these labels into CCR dimensional values was made on the basis of their definitions and on accompanying examples provided by the manuals. For every dimension, we determined how that dimension can be mapped on the relational inventory of the frameworks, and where the problems come from when the dimension is not easily translatable between frameworks. Appendix C presents the dimensions and features included in the UniDim approach; Appendix D shows the correspondences between RST-DT and PDTB based on the decompositions. Here, we give a brief summary of how well dimensions could be applied to the labels in different frameworks. For a detailed discussion of any problematic mappings, see Sanders et al. (2018).

**Polarity** Most of the classes proposed by PDTB and RST-DT distinguish relations with a `positive` value for polarity from labels with a `negative` value. Determining polarity was therefore fairly straightforward. For example, all labels belonging to PDTB's class Temporal are `positive`, and all labels belonging to Comparison are `negative`. However, both PDTB and RST-DT include a few relation labels that seem to apply to both `positive` and `negative` relations. One example is PDTB's Conjunction relation, that can be signaled by *and* (a `positive` connective) and *but* (a `negative` connective). The polarity for this relation label is therefore underspecified. Similarly, RST-DT's Condition can be marked by *if* (`positive`) and *unless* (marking `negative` conditionals). However, for explicit relations from these underspecified categories, the polarity can easily be determined on the basis of the connective used to signal the relation. Polarity can therefore be successfully mapped to PDTB and RST-DT in most cases.

**Basic operation** PDTB and RST-DT generally present clusters of relations along the basic operation dimension. For example, PDTB's class Contingency only contains `causal` and `conditional` relations. However, both frameworks also include a few labels that cannot be mapped as easily. For RST-DT, a problematic label is Antithesis: examples in the manual indicate that these relations can be `negative additive` as well as `negative causal`. Again, in most underspecified cases, connectives can indicate which basic operation applies to a specific relation. However, certain connectives are themselves underspecified: consider *but*, which can be used to mark `causal` and `additive` relations. We therefore argue that most, but not all, relations can be specified for their basic operation.

**Source of coherence** Mapping source of coherence to PDTB and RST-DT relations is problematic, and quite a few labels remain underspecified. Certain labels are clearly defined as `objective` or `subjective` (e.g., RST-DT's Evaluation or PDTB's Justification are clearly `subjective`), but for approximately one-fifth of the labels, the source of coherence cannot be determined. Given that connectives typically do not signal subjectivity in English (unlike in other languages such as Dutch and Chinese), manual annotation would be required to classify specific instances of these relations in terms of source of coherence.

**Implication order** The mapping of this dimension onto PDTB and RST-DT relation labels is not straightforward due to different notions of order in the frameworks. CCR looks at the event order of the segments. If the first segment of a relation presents a cause, for example, then the relation has a `basic` order. If it presents a consequence, then the relation has a `non-basic` order. By contrast, PDTB and RST-DT do not determine argument order based on their occurrence in the text, but rather on the syntactic attachment to the connective (PDTB) and the nucleus-satellite order (RST-DT). Consequently, PDTB and RST-DT distinguish between similar relations with opposite orders in their relation inventory (e.g., PDTB Reason and Result). To illustrate the difference, consider the following examples:

(13) Toby likes to listen to recordings of ocean waves because ocean sounds soothe him.

(14) Because ocean sounds soothe him, Toby likes to listen to recordings of ocean waves.

These relations would receive different values for implication order in CCR: (13) has a `non-basic` order, and (14) has a `basic` order. In PDTB, however, these examples would receive the same label: Arg2 presents the reason in both relations, so they would be labeled as Reason. Similarly, in RST-DT, both *because*-clauses are considered the satellite, and so both examples would receive the label Reason. The notion of order in the PDTB and RST-DT frameworks therefore focuses more on the relative importance of the segments – which segment is central in the discourse. Both types of order (implication and relative importance) are valid and justifiable. In order to be able to accurately represent the order of such relations, we would therefore need to take into account the implication order and the textual Arg1-Arg2 order (PDTB) or nucleus-satellite order (RST-DT).

Many of PDTB's and RST-DT's labels remain underspecified for implication order because of the different notions. In practice, mapping the order of explicit instances between frameworks should be possible by exploiting the information regarding the connective and the Arg1-Arg2 or nucleus-satellite order. For example, a Reason relation marked by *because* with an Arg2-Arg1 order can be mapped to a `basic` implication order.

**Temporality** Both PDTB and RST-DT set temporal relations apart from other types of relations, and distinguish different types of temporal relations. PDTB distinguishes between SYNCHRONOUS and ASYNCHRONOUS relations; RST-DT distinguishes between TEMPORAL-SAME-TIME vs. TEMPORAL-BEFORE and TEMPORAL-AFTER relations. Only a few labels seem to apply to both `temporal` and `non-temporal` relations, namely RST-DT's BACKGROUND and CIRCUMSTANCE labels. In the mapping, the value for temporality is therefore underspecified for these labels.

**Additional features** After decomposing all PDTB and RST-DT labels using the five dimensions, we evaluated which additional features were needed in order to capture more fine-grained differences that could not be represented by the CCR dimensions. In doing so, we ensured that most end labels could be classified by a unique combination of features. We took a bottom-up approach for selecting the features, meaning that we used features that were already present in the annotation frameworks at hand. In the future, if new frameworks are analyzed according to UniDim and the need for additional features arises, the set can easily be expanded.

Within the class of additives, the new features that were included were:

- *Specificity*, used to distinguish relations where one argument is more specific than another (such as PDTB SPECIFICATION and GENERALIZATION),
- *Lists*, used to distinguish LIST relations from non-list relations, and
- *Alternatives*, used to distinguish additive relations in which the two segments are presented as alternatives (e.g., PDTB's CHOSEN ALTERNATIVE).

Within the class of causals, the new features were:

- *Goal-orientedness*, used to distinguish relations where one of the segments concerns an intentional, goal-directed action by an agent (these relations are typically marked by *in order to*), and
- *Conditionality*, used to distinguish relations for which the cause is not yet realized.

## 3.4 Discussion and conclusion

In order to make more progress in the study of discourse coherence, researchers need to be able to make optimal use of existing discourse-annotated data. The various frameworks that have been used to annotate the data differ from each other in the relations that they distinguish, the labels they give relation types, and the underlying definitions of labels. As a result, the interoperability of the corpora is blocked.

To address this issue, several proposals have been made to unify frameworks. These proposals can roughly be categorized into two types: proposals using an intermediate framework and proposals using an intermediate representation. The first type, using an intermediate framework, generally focuses on creating a framework of "core" coherence relations. Hovy & Maier (1995) proposed a taxonomy of roughly 70 discourse relations, Bunt & Prasad (2016) proposed a set of 20 relations, and Benamara & Taboada (2015) a hierachy of 26 relations. These efforts can function as useful tools for future annotation, but since mapping different frameworks was not their primary goal, problems arise when the frameworks differ in granularity (i.e., when one framework distinguishes a label that another framework does not). Intermediate frameworks cannot easily provide a one-to-many mapping in case there is no direct corresponding label in one of the frameworks. The existing data would therefore have to be re-annotated according to a new framework in order to be fully interoperable.

Another solution for unifying frameworks is the second type of proposals: using an intermediate representation. OLiA is one example of such a proposal. The intermediate representation is designed as a hierarchical structure of *subclasses*. In order to map a relational label from one framework, one can simply find the corresponding superclass in the other framework. However, in cases where the corresponding superclass occurs very high in the hierarchy, the label is mapped to a very general class (e.g., all additive labels in another framework). This could create difficulties for the mapping in practice.

A second proposal using an intermediate representation is the Unifying Dimensions approach. All labels from PDTB and RST-DT were decomposed according to their values on several unifying dimensions. These dimensions allow for clusterings across frameworks (i.e., cluster adversative relations from non-adversative relations). Researchers can use UniDim to "translate" labels from one framework to another. Additionally, the UniDim approach is useful for pinpointing the exact differences and similarities between relation types proposed by different approaches.

No mapping can overcome the differences in granularity that exist between frameworks (i.e., a mapping scheme cannot add information that is absent in one of the corpora). However, the UniDim mapping approach provides a solution for this issue by allowing for all relational labels to be mapped based on their similarity. In cases where one framework distinguishes a type of relation that another framework does not, the interlingua indicates which labels in the "underspecified" framework are the most likely counterparts (given their similar specification on other features). The interlingua also gives insight into where the differences lie, which is something that an intermediate relation label framework cannot do. As a result, every relational label from one framework can be mapped to the closest corresponding match in another framework, even when no exact corresponding label exists.

In conclusion, we have seen how PDTB 2.0 and RST-DT differ from each other and how this complicates interoperability. Several efforts to unify frameworks have been discussed, ending with a new, feature-based approach referred to as UniDim. This approach is seen as a promising step forward in increasing interoperability between different frameworks, corpora, and paradigmata. The decomposition of relation labels using UniDim was based on the definitions that the frameworks provided in their manuals. However, the annotations might differ in practice, so the approach still needs to be validated. In the next chapter, we evaluate the proposed mappings using annotated data from several different corpora.

# Chapter 4

# Mapping data from different discourse frameworks[1]

In the previous chapter, the lack of interoperability between discourse frameworks was established, and the Unifying Dimensions proposal was introduced to offer a solution for this issue. UniDim's mapping is based on the definitions and examples provided in the annotation guidelines. Ideally, the theoretical explanations provided in the guidelines correspond to the actual annotations in the corpora, and the theoretical mapping would therefore also correspond to the mapping in practice. In order to investigate this and validate the theoretical mapping, we compared annotations of different frameworks on the same data, to ensure that the provided annotations correspond to each other.

This effort has been undertaken for two combinations of frameworks: (i) we annotated a corpus of spoken data using both PDTB and CCR (as described in Rehbein, Scholman & Demberg, 2016), which allowed for a comparison of the two annotation layers; and (ii) we compared the annotations of 385 newspaper articles of the Wall Street Journal that are contained in both the PDTB and RST-DT (as described in Demberg, Asr & Scholman, submitted). In this chapter, both efforts and their results are discussed. The systematic clusters of disagreement that are identified in this chapter function as input for the experiments reported in the remainder of this thesis.

In addition to evaluating the Unifying Dimensions approach, the work that is reported in this chapter contributes to the field in several ways. For the first study,

[1] The PDTB–CCR mapping that is discussed in this chapter was reported in Rehbein, I., Scholman, M.C.J., and Demberg, V. (2016). Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 16)* (pp.23–28). Portoroz, Slovenia.
The PDTB–RST-DT mapping was reported in Demberg, V., Asr, F.T. and Scholman, M.C.J. (submitted). How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Submitted to Dialogue & Discourse.* Some parts of this chapter are identical with these publications.

we annotated a corpus of spoken data. The amount of spoken data annotated with coherence relations was limited. By increasing this amount, another step is made towards being able to use this type of data for domain adaptation in discourse relation classifiers. Second, the data that was annotated for the first study contributes to the body of data available that carries annotations from multiple frameworks. This type of data can be used to improve understanding of differences and similarities between frameworks. The second study compares existing annotations done using different frameworks on the same data. These two studies present the first efforts investigating the compatibility of these annotations.

## 4.1 PDTB vs. CCR

Despite the existence of many different frameworks and their corresponding annotated corpora, few resources exist with annotated labels from multiple frameworks. Moreover, most discourse-annotated resources are based on written, rather than spoken, text. In order to increase the accuracy of discourse relation classifiers for genres other than newspaper articles, more data is needed from different sources (this issue is referred to as domain adaptation). This is why we annotated a corpus of spoken data using both PDTB 3.0 and CCR (Rehbein, Scholman & Demberg, 2016). The annotations that make up the resulting corpus, referred to as Disco-SPICE, were then used to verify the proposed mapping from PDTB labels to CCR dimension values.

The creation of a spoken English corpus with annotations from the PDTB provides researchers with a means to compare written and spoken text: the Disco-SPICE corpus is interoperable with existing PDTB resources, which allows for an evaluation of distributional differences between the domains. It had already been established in previous works that written and spoken texts are produced and processed differently (Biber, 1991; Chafe & Tannen, 1987; Horowitz & Samuels, 1987; Tannen, 1982): spoken discourse is characterized by a high degree of interactivity, shorter sentence length on average, and a higher proportion of disfluent sentences. Less was known regarding distributional differences of coherence relations. Using the Disco-SPICE annotations, we were able to show that the use of coherence relations in the spoken domain does in fact differ from the written domain; for example, relations are more often explicit in spoken data compared to written data, and CONCLUSION and RESTATEMENT relations are more common in spoken text (see Rehbein et al., 2016). In this remainder of this section, we will focus on the evaluation of the mapping between PDTB and CCR. The differences between written and spoken discourse do not affect this mapping, and will not be considered further.

### Method

**Data** The data consisted of spoken text from the SPICE-Ireland corpus (Kallen & Kirk, 2012); a corpus of spoken Irish English containing a variety of genres. We

| | *Genre* | | |
|---|---|---|---|
| | **Broadcast** | **Telephone** | **Total** |
| Total nr. of sentences | 1,507 | 2,717 | 4,224 |
| Total nr. of words | 20,801 | 20,239 | 41,040 |
| Total nr. of PDTB relations | 1,244 | 1,201 | 2,445 |
| Total nr. of CCR relations | 1,064 | 1,005 | 2,069 |
| Nr. of explicit rels per 100 sent. | 44.0 | 25.5 | 34.75 |
| Nr. of implicit rels per 100 sent. | 27.2 | 12.6 | 19.9 |

**Table 4.1:** Descriptive statistics of the corpus per genre and overall. The number of explicit and implicit coherence relations is represented as the average frequencies per 100 sentences/speech units.

annotated 10 SPICE-Ireland texts consisting of broadcast interviews (public) and 10 texts consisting of private telephone conversations.

**Annotation procedure** Two trained coders annotated all implicit and explicit coherence relations according to the PDTB 3.0 (see Appendix A) and CCR framework. All texts were annotated by one coder following the PDTB guidelines, and by the second coder following the CCR guidelines. The CCR annotations adhered to the four original dimensions (polarity, basic operation, source of coherence and order). The additional features proposed in UniDim were not included in the annotation scheme. Note that in the version of CCR used for this experiment, temporality was conflated with the basic operation dimension (i.e., `temporal` was annotated as an additional possible value of *basic operation*, rather than as a separate dimension, cf. Scholman et al., 2016).

Table 4.1 shows the size of the subcorpora from the different genres. In total, 2,445 coherence relations were annotated according to the PDTB 3.0 framework, and 2,069 coherence relations were decomposed into CCR dimensions. The lower number of CCR dimensions can be attributed to theoretical differences between the two frameworks. Consider, for example, relations with multiple connectives such as *but then*: in PDTB, each connective receives a label, while in CCR, those cases obtain only one label. Another example is instances that receive the PDTB label EntRel or NoRel. These instances are not annotated in CCR.

**Inter-annotator agreement (IAA)** A subset of the data was annotated by both annotators using the same framework in order to determine inter-annotator agreement. For CCR, this amounts to 89 explicit coherence relations and for the PDTB framework, annotations for 175 explicit relations were compared.[2] In order to evaluate the agreement on these relations, percentages of agreement are calculated, as

[2] The difference in the number of relations annotated using PDTB and CCR was due to time restrictions.

well as Kappa values and AC$_1$ scores. Before we move on to a discussion of the inter-annotator agreement in Disco-SPICE, a brief explanation of these measures is given to provide a background against which the results can be evaluated. For a more detailed discussion of calculating inter-annotator agreement for discourse-annotated data, we refer readers to Hoek & Scholman (2017) and Spooren & Degand (2010).

The simplest measure of agreement between annotators is the *observed agreement*, which is the percentage of agreement. However, this measure is often not suitable for calculating reliability, because it does not take into account *chance agreement* (Scott, 1959). Chance agreement occurs when one or both annotators rate an item at random. This type of agreement should not contribute to a measure of inter-annotator reliability because it can inflate the overall agreement (Artstein & Poesio, 2008). In order to get a reliable agreement score, observed agreement has to be adjusted for chance agreement. The proportion of chance agreement must be estimated, because we cannot know when annotators agreed by chance (Gwet, 2001). Cohen's Kappa and AC$_1$ both correct the agreement scores for chance agreement.

Cohen's Kappa is the most commonly used inter-annotator agreement measure, but it is known to present some problems in particular situations; specifically, Kappa's values are sometimes relatively low, despite a high percentage of observed agreement. This problem is known as the "Kappa paradox" (Feinstein & Cicchetti, 1990). This paradox occurs because Kappa is sensitive to certain characteristics of data that are very typical of discourse data: frameworks often distinguish between a large number of categories, and datasets are characterized by an uneven distribution of these categories (that is, some types of relations occur more frequently than others). AC$_1$ was introduced to address these issues, and therefore seems to be more suitable for discourse-annotated data. AC$_1$ has been applied often in the medical field (e.g., Bryant et al., 2013; Crowle et al., 2017; Fuller et al., 2017; Marks et al., 2016) and has also been used in the computational linguistics field (Besser & Alexandersson, 2007; Haley, 2009; Hillard et al., 2007; Kranstedt et al., 2006; Purpura & Hillard, 2006; Yang et al., 2006), but little research in the field of discourse annotation has used AC$_1$ as of yet (but see Hoek & Scholman, 2017). We therefore report both Kappa and AC$_1$ in this section.

Table 4.2 shows the annotators' agreement on annotations using CRR dimensions and PDTB labels for these subsets.[3] For CCR, the agreement scores are reported overall and per dimension. Calculating agreement separately for every dimension provides a clear overview of where exactly confusions or disagreements arise (see also Hoek, 2018, for a discussion on calculating agreement when using CCR).

For CCR, the Kappa ranges between .50 and .77 for the individual dimensions and AC$_1$ ranges between .63 and .89. The higher AC$_1$ scores for every dimension can be attributed to the uneven distribution of the values; for example, `positive` relations are more prevalent than `negative` relations. AC$_1$ takes this uneven distribution into

[3] The agreement scores reported in this chapter were calculated using the R package agree.coeff2.r.

| | *CCR* | | | *PDTB 3.0* | | |
|---|---|---|---|---|---|---|
| **Dimension** | **% agr.** | $\kappa$ | **AC$_1$** | **% agr.** | $\kappa$ | **AC$_1$** |
| Polarity | 89.8 | .75 | .83 | - | - | - |
| Basic operation | 91.0 | .77 | .89 | - | - | - |
| Source of coherence | 78.7 | .50 | .63 | - | - | - |
| Order | 87.6 | .70 | .84 | - | - | - |
| All | 67.4 | .60 | .65 | 84.6 | .79 | .83 |

**Table 4.2:** Inter-annotator agreement (percentage agreement, Kappa scores, and AC$_1$ scores) for coherence relations in Disco-SPICE carrying CCR and PDTB annotations.

account and adjusts the amount of estimated chance agreement (see also Hoek & Scholman, 2017).

The decomposition of relations into dimensions reveals on what aspect of the relation annotators specifically disagree: the source of coherence dimension shows the lowest agreement score. As noted in the previous chapter, this result has been found in other studies as well (e.g. Sanders et al., 1992; Scholman et al., 2016). The agreement on CCR's combined dimensions (exact match: an annotation counts as correct if the coders assigned the same values for all four dimensions) is 67%, with a Kappa of .6 and AC$_1$ of .65. These relatively low scores can largely be attributed to the annotations of source of coherence.

For PDTB, agreement on explicit annotations corresponds to 84.6%, with a Kappa of .79 and AC$_1$ of .83. The different values for overall agreement between CCR and PDTB cannot be compared because they are based on different annotation frameworks, different items and a different amount of items (see also Hoek & Scholman, 2017). In other words, we cannot be sure whether the lower score for CCR is because the annotators are more trained in annotating PDTB labels, because the items in one subset were more ambiguous than in the other subset, or because the higher number of observations in PDTB affected the scores. What we can conclude is that the agreement scores for both CCR (with the exception of source of coherence) and PDTB are deemed acceptable for coherence relation annotations (cf. Spooren & Degand, 2010).

### Results

All PDTB annotations and CCR annotations for relations in the SPICE-Ireland corpus were mapped onto each other. The correspondences were then evaluated using the Unifying Dimensions mapping. This decomposition allows us to test whether coherence relations annotated according to the PDTB framework by one coder fall into corresponding categories when annotated by the other coder using CCR. The

| **CCR ↓** | | **PDTB→** | | TEMP. | | CONT. | | | | | COMP. | | EXP. | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Polarity | Basic operation | Source of coherence | Order | Synchronous | Asynchronous | Cause | Cause_belief | Cause_speechact | Condition | Cond._speechact | Concession | Contrast | Disjunction | Substitution | Conjunction | Equivalence | Instantiation | Specification |
| pos | temp | obj | na | **68** | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| pos | temp | obj | bas | 13 | **67** | 2 | 0 | 0 | 1 | 0 | 2 | 1 | 0 | 0 | 16 | 0 | 0 | 0 |
| pos | temp | obj | n-b | 4 | **9** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 1 |
| pos | caus | obj | bas | 0 | 3 | **17** | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 |
| pos | caus | obj | n-b | 0 | 0 | **12** | 5 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 6 |
| pos | caus | subj | bas | 0 | 1 | **25** | **40** | **53** | 0 | 0 | 4 | 0 | 0 | 0 | 6 | 4 | 0 | 6 |
| pos | caus | subj | n-b | 0 | 2 | **36** | **32** | **47** | 1 | 0 | 0 | 1 | 0 | 0 | 2 | 32 | 24 | 23 |
| pos | cond | obj | und | 0 | 1 | 1 | 0 | 0 | **30** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pos | cond | subj | und | 0 | 0 | 0 | 0 | 0 | **58** | **93** | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| neg | caus | obj | und | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 2 | 0 | 7 | 0 | 0 | 0 | 0 |
| neg | caus | subj | und | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **25** | 10 | 0 | 0 | 0 | 0 | 0 | 1 |
| neg | add | obj | na | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 20 | **33** | **25** | **22** | 4 | 4 | 3 | 3 |
| neg | add | subj | na | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 32 | **43** | **55** | **64** | 6 | 2 | 3 | 4 |
| pos | add | obj | na | 8 | 6 | 3 | 9 | 0 | 1 | 0 | 2 | 2 | 0 | 7 | **31** | **14** | **16** | **24** |
| pos | add | subj | na | 5 | 3 | 3 | 9 | 0 | 5 | 7 | 3 | 3 | 20 | 0 | **31** | **42** | **50** | **30** |
| | | | Count | 53 | 105 | 300 | 22 | 15 | 77 | 14 | 56 | 206 | 20 | 14 | 538 | 45 | 38 | 143 |

**Table 4.3:** Distribution (%) of explicit and implicit relations, only labels and categories where $n$ >10 (und: underspecified, na: not applicable) and raw counts.

mapping resulted in the correspondence matrix shown in Table 4.3.[4] The cells with the bold-font, underlined numbers correspond to the predicted mapping as proposed in the Unifying Dimensions approach.

Several issues that were discussed in the previous chapter can be identified in the mapped data in Table 4.3. For example, we can see that most relation labels in PDTB are underspecified regarding their source of coherence. The same goes for CCR's order. Similarly, it can be noted that many labels in the EXPANSION class fall into the same CCR categories: `positive, additive, objective/subjective`. The additional features that were suggested in the Unifying Dimensions proposal (but not yet included in this study) should capture these more fine-grained distinctions.

PDTB 3.0 distinguishes causal relations with a speechact reading (i.e., Cause_speechact and Condition_speechact), whereas in CCR, speechact relations are classified as

[4] For a complete version of the table, including PDTB level-three distinctions, see `http://www.sfb1102.uni-saarland.de/?page_id=2582`.

subjective relations. An additional distinction in CCR's source of coherence would account for these speechact relations.

In an ideal situation, every type of relation distinguished in PDTB would be annotated as a similar type in CCR; for example, all relations with the PDTB label Cause should be annotated as `positive cause` relations according to the CCR dimensions. Table 4.3 reveals that this is not always the case. Overall, 70% of the PDTB relations were categorized as belonging to the target CCR class. It is not immediately clear what the cause is for mismatches: they could be due to different interpretations by the annotators (as reflected in inter-annotator disagreement), or they might be caused by more theoretical differences between the frameworks regarding the annotation process and coherence relation definitions. To determine the common causes of mismatches in the data, we analysed a random sample of 50 disagreements.

19 of the 50 disagreements can be attributed to annotation errors; that is, after discussing the relation, one coder agreed that another label was more appropriate. An additional 7 disagreements were caused by differences in interpretation: the coders had annotated different relations with similar spans and could agree on both labels according to a different interpretation. The remaining 24 cases of the sample were caused by differences in operationalization of the frameworks.

Most of the mismatches due to annotation operationalizations can be classified as belonging to one of three categories. The first category of disagreements concerns Contrast and Concession relations: among these `negative` relations, coders often disagreed on the `causal` vs. `additive` basic operation (52% of relations annotated as Concession according to PDTB). Note that distinguishing between contrastive and concessive coherence relations is a well-attested difficulty (see, for example, Robaldo & Miltsakaki, 2014; Zufferey & Degand, 2013). PDTB and CCR seem to have different operationalizations to make this distinction. In PDTB's Concession relations, one argument creates an expectation that the other denies. What exactly constitutes an expectation is up to the coder. In CCR's `negative causal` relations, P and the negative counterpart of Q function (i.e., a denied cause of consequence). Coders can make use of a paraphrase test to determine whether a negative relation is additive or causal: reverse the polarity and then determine the basic operation. This is illustrated in Example (15), which is taken from a fragment where a speaker mentions wanting to go to a party, but needing to work in a shop instead. Following PDTB practices, the first argument is printed in italics, and the second in bold. From the PDTB perspective, the first argument leads to an expectation that the speaker will attend the party, which is then denied in the second argument. This licenses a Concession.contra-expectation relation. Crucially, in CCR, this example would not be classified as `negative causal`. Even though the paraphrase is acceptable with *because*, it does not have the same meaning as the original relation: the second argument of the relation is not the true cause or result of the first argument. The paraphrase test therefore results in a stricter definition of what can be considered a `negative causal` relation, which in turn causes disagreement between the frameworks.

(15) Original:
*Cos I wouldn't mind going down.* <u>However</u> **the shop won't be closed I'd say until about seven.** (And that's a bit too late really to go to Derry.)

Paraphrase:
*Cos I wouldn't mind going down.* <u>So/Because</u> **the shop *will* be closed around seven.**

The second category of disagreements concerns argumentative relations. Many relations annotated as PDTB Instantiation, Specification and Equivalence (all belonging to the EXPANSION class) were annotated as `subjective causal` relations in CCR. In CCR, if an implicit relation can be paraphrased as "one segment provides an argument for a claim made in the other segment," and this relation can be marked by the connective *because*, the relation is classified as a `subjective causal` relation. In PDTB, however, annotators are free to insert a connective of choice without a prescribed order of insertion. This can lead to different interpretations of the same relation. Example (16), from a telephone conversation discussing the TV programme, illustrates this issue: for the PDTB label, the second segment (printed in bold) was interpreted as specifying why the TV programme is brilliant. This reading can be made explicit by inserting the connective *specifically.* In CCR, however, the second segment was interpreted as an argument for the claim that the TV programme is brilliant; a reading that can be marked by mentally inserting the connective *because.*

(16) Speaker A: Have you seen the TV tonight?
Speaker B: No what's on?
Speaker A: *It's brilliant.* **There's Sex and the City and uhm is Ally McBeal on or** ... No I can't remember. There's something really a a really good line-up.

a. PDTB: EXPANSION.Specification
b. CCR: `positive causal subjective backward`

The third category of disagreements – albeit less frequent than the previous two categories – concerns `negative additive` relations. CCR makes use of a substitution test to determine polarity: if a relation's arguments can be connected by *but*, the relation is coded as `negative`. In PDTB, however, relations that can be connected by *but* can also fall in the EXPANSION class. Example (17) illustrates this issue. The fragment discusses the heritage of a horse. The first argument mentions the horse's father (the Northern-based sire). The PDTB label Specification was assigned because the second argument specifies the whereabouts of the sire. In CCR, however, this relation was annotated as `negative`, because *but* can be inserted between the segments. This reading is similar to that of a qualification relation: S1 seems to leave open the possibility that the horse's father is still alive, but this is denied in S2. The second argument therefore modifies the first by canceling the strongest possible interpretation of the first argument (see Hoek, 2018; Pander Maat, 1998). The relation is also annotated as `subjective`; not because the second argument is in fact `subjective`, but because the uttering of the second argument to cancel the strongest possible interpretation of the first argument creates a `subjective` link between the segments in this context.

(17) She's a bay mare fifteen two hands high so she's not very big but big enough for jumping. She's – black mane and tail of course. *She's by a Northern-based sire.* **I think he's dead now perhaps.** Wren's Hill – and out of a chasing-bred mare.

a. PDTB: EXPANSION.Specification

b. CCR: `negative additive subjective`

The main cause of mismatches falling into these last two categories is hence the operationalization in terms of which connective should be inserted first for implicit relations: CCR relies on a limited set of connectives in a specific order of importance to determine the values for the dimensions (with *but* and *because* being the strongest and *and* being the weakest indicator), whereas PDTB does not categorize connectives in a similar way to determine the end labels.

### Discussion PDTB–CCR mapping

In sum, the mapping of PDTB and CCR annotations has shown that PDTB relations fall into the target CCR class relatively often (70% of the cases). These results indicate that the mapping of PDTB labels is successful, specifically in terms of UniDim's polarity and basic operation dimensions. Most mismatches did not occur due to wrong decomposition of the labels; rather, they seem to be due to (a) annotator errors, irrespective of the frameworks, (b) differences in the definition of what constitutes a `negative causal` relation, or (c) the operationalization of the annotation process leading to different annotations.

Crucially, this study shows that disagreements in mapped data can be caused by the systematic differences in the operationalizations of the annotation schemes; that is, differences in how exactly the coders are supposed to decide between coherence relations. This is an important characteristic of mapped data to keep in mind when interpreting the results.

The current study also has several limitations that need to be considered. First, the annotations using different frameworks were done by annotators that trained together. This differs from mapping efforts where the annotators of different approaches will not have trained together. It is possible that the shared training in the current study led to more inter-framework agreement compared to when annotations are mapped from independent annotators. Second, the relations included in the Disco-SPICE corpus were first segmented, and then annotated using different frameworks. This differs from mapping efforts between independent annotations, where the data is segmented using different segmentation rules. Different segmentations might lead to different interpretations and therefore less complete mappings (or less agreement).

Moreover, the frameworks that were used were originally designed for written discourse, not spoken discourse. This means that some relational categories in the frameworks might not be entirely applicable to spoken discourse, and some categories that are relevant in spoken discourse might have been missing. Finally, the results from the current effort are only based on one mapping between two frameworks. In order to know whether these results can be generalized, we need to verify these results with another mapping between different frameworks.

The next section presents another mapping effort that addresses these issues: data from overlapping parts of the original PDTB 2.0 and RST-DT corpora were mapped onto each other in order to evaluate the compatibility of the annotations and the accuracy of the Unifying Dimensions approach.

## 4.2 PDTB vs. RST-DT

The PDTB and RST-DT corpora partly consist of the same texts, thereby allowing for a large-scale comparison of annotations. In order to evaluate the compatibility of the annotations, we mapped them onto each other (as reported in Demberg, Asr & Scholman, submitted). This mapping differs from the PDTB–CCR mapping discussed in the previous section because the data was annotated by the original coders from the two frameworks. This has implications for the segmentation of the relations, which in turn has implications for the mapping, as we will see in the next section.

### Method

**Data** PDTB 2.0 and RST-DT annotations overlap for 385 newspaper articles of the Wall Street Journal corpus. The mapping is based on this intersection of the PDTB and RST-DT.

**Alignment** A first challenge lies in aligning the data: PDTB and RST-DT apply different segmentation principles for identifying the arguments or elementary discourse units of relations. First, the two frameworks differ in what they consider an argument (PDTB terminology) or an "elementary discourse unit" (RST-DT terminology). Note that in the remainder of this chapter, "segment" is used to refer to the text elements that make up the relation (i.e., the argument, EDU, nucleus or satellite). Second, the frameworks differ in the resulting discourse structure: PDTB only annotates relations within and between adjacent sentences, and relations marked by explicit connectives. RST-DT, on the other hand, annotates discourse trees spanning

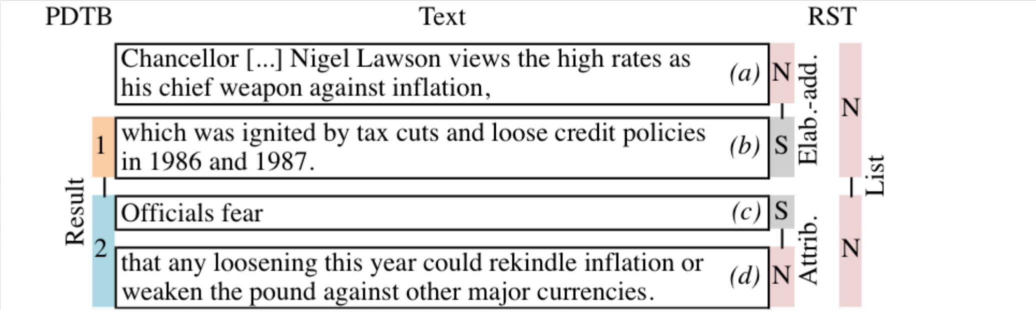

PDTB
Text
RST
Chancellor [...] Nigel Lawson views the high rates as his chief weapon against inflation, (a) N
which was ignited by tax cuts and loose credit policies in 1986 and 1987. (b) S
Officials fear (c) S
that any loosening this year could rekindle inflation or weaken the pound against other major currencies. (d) N
1
2
Result
Elab.-add.
Attrib.
N
N
List

**Figure 4.1:** PDTB and RST-DT annotations for a paragraph of wsj_1172. *1* refers to Arg1 in PDTB; *2* refers to Arg2. *N* refers to Nucleus in RST-DT; *S* refers to satellite. (a-d) refer to RST-DT's EDU's.

the entire document. This results in a higher number of RST-DT relations compared to PDTB relations.

For the automatic alignment procedure, we therefore used PDTB as a starting point, and aimed at identifying the corresponding RST-DT label for every PDTB relation. This was done by finding the RST-DT segments that were most similar (and not necessarily identical) to the PDTB segments. Although we aimed to map as many coherence relation labels as possible, we were careful to map only those labels where annotators inferred the same relation – if the RST-DT annotators annotated a relation holding between two segments, and the PDTB annotators marked a relation between two slightly different segments that led to a different interpretation, these labels should *not* be mapped.

To ensure that valid alignments can be identified even when the relational segments differ slightly, several flags were implemented in the alignment procedure that marked instances for which the mapping was potentially problematic. One of these flags relied on RST-DT's Strong Nuclearity hypothesis, which states that when a relation is postulated to hold between two spans of text, it should also hold between the nuclei of these spans (Marcu, 2000). To illustrate this, consider the PDTB Result relation in Figure 4.1, which presents the PDTB (left) and RST-DT (right) annotations for a fragment of a Wall Street Journal article. The nucleus path of the first segment of RST-DT's List relation leads to segment (a), not to segment (b) (which corresponds to PDTB's Arg1). Following the Strong Nuclearity hypothesis, PDTB's Result relation should not map onto RST-DT's List relation, and these labels should therefore not be aligned.

### Results of alignment procedure

In total, 76% of PDTB annotations from the joint corpus (a total of 5141 relations) were mapped to a corresponding RST-DT annotation. 52% of these instances have

[4] The nucleus path is the path from a complex segment in a high-level relation to the nucleus in the lowest-level relation.

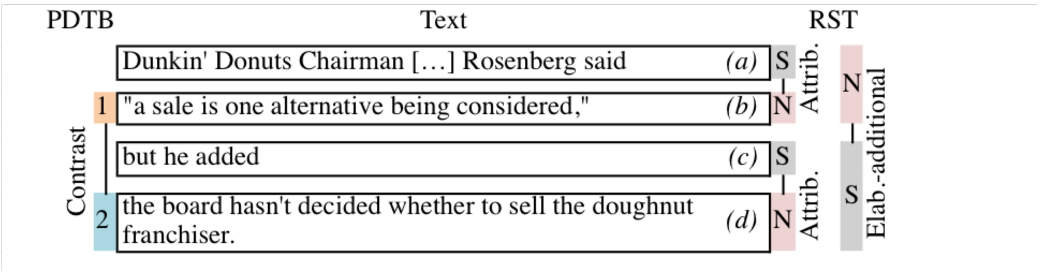

PDTB
Text
RST
Dunkin' Donuts Chairman [...] Rosenberg said (a)
"a sale is one alternative being considered," (b)
but he added (c)
the board hasn't decided whether to sell the doughnut franchiser. (d)
1
2
Contrast
S
N
S
N
Attrib.
Attrib.
N
S
Elab.-additional

**Figure 4.2:** PDTB and RST-DT annotations for a paragraph of wsj_1176. Note: only the PDTB annotation that is relevant for this example is included.

directly corresponding segment spans. The remaining 48% of data included in the mapping analysis consists of relations for which at least one of the RST-DT segments was more complex (i.e., contains multiple EDU's) than the PDTB segment. In order to get more insight into whether these more complex relations are aligned correctly, we randomly selected a subset of 100 instances and evaluated if the algorithm was justified in mapping the labels.

Manual analysis showed that 95 relations were mapped successfully, while the mapping of 5 instances was unjustified. In these five cases, the nucleus of a larger RST-DT span matched PDTB's argument, but the annotators did not evaluate the same type of relation. To illustrate this, consider the example in Figure 4.2. PDTB coders assigned a CONTRAST label between segments (b) and (d), whereas RST-DT coders assigned an ELABORATION-ADDITIONAL label between segments (a-b) and (c-d). The inclusion of the two attributions changes the interpretation of the corresponding relation: the relation focuses on the speaker saying something and then adding to that (expressed by an ELAB.-ADDITIONAL relation), instead of on the content of what the speaker is saying (expressed by a CONTRAST relation). ATTRIBUTION can therefore (in a subset of cases) "block" another interpretation.

To quantify the problem that intervening attributions might present, we counted the occurrence of intervening ATTRIBUTION relations. In total, 595 relations (12% of the data) had at least one ATTRIBUTION relation in one of the segments; 49 instances (less than 1% of the data) had two or more intervening attributions (as in Example 4.2). The problem of attribution leading to different interpretations will only apply to some relations from this set of 49 instances. We therefore decided to include instances with attributions in our analysis.

24% of PDTB relations (1621 instances) were flagged as possibly problematic. That is, there is a higher chance that the segments that PDTB and RST-DT identified do not correspond to each other. These instances were therefore excluded from the analysis. To assess whether the alignment was correct in excluding these relations, we again randomly selected and manually evaluated 100 instances.

We found that 76 items were correctly flagged as unjustified mappings. The 24 remaining relations do actually represent valid mappings. In 13 of these cases, the

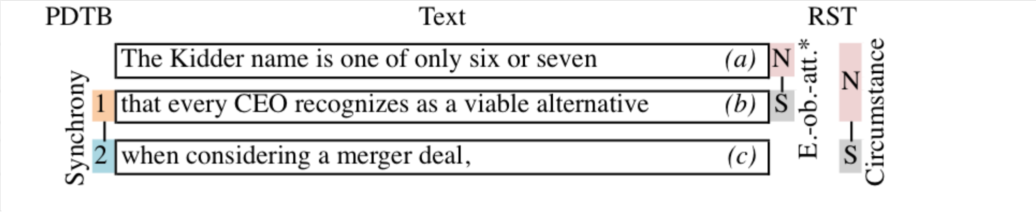

PDTB
Text
RST
The Kidder name is one of only six or seven (a)
that every CEO recognizes as a viable alternative (b)
when considering a merger deal, (c)
Synchrony
1
2
N
S
E.-ob.-att.*
N
S
Circumstance

**Figure 4.3:** PDTB and RST-DT annotations for a paragraph of wsj_0604. Note: only the PDTB annotation that is relevant for this example is included. *The full label is Elaboration-object-attribute.

RST-DT annotation was inconsistent with the Strong Nuclearity principle; that is, the Strong Nuclearity principle was violated but the annotations did match. This is illustrated in Figure 4.3. PDTB annotated a Synchrony relation between segment (b) and (c). RST-DT annotated a Circumstance relation between segment (a-b) and (c), with segment (a) being the nucleus. Consequently, the nucleus path could not be traced to PDTB's Arg1, and the relation was flagged. However, the Circumstance relation does not hold between *The Kidder name is one of only six or seven* and *when considering a merger deal*; it holds between what PDTB marks as Arg1 and Arg2. The Strong Nuclearity hypothesis therefore does not hold in this case.

Another typical case for automatic flagging was caused by PDTB's annotation constraint for annotating only adjacent implicit relations. When two adjacent sentences convey a similar message and are followed by a third sentence which is Arg2, the PDTB and RST-DT mapping was incorrectly flagged, because in those cases, the second but not the first sentence has to be annotated as Arg1 in PDTB.

### Correspondence between mapped labels

After aligning the two annotation layers, the correspondence between the labels was evaluated. The mapping showed that agreement on explicit relations is reasonable, whereas agreement between the frameworks on implicit relations is low. Each is discussed in turn.

**Explicit relations.** Table 4.4 shows the mapping of PDTB level-2 annotations onto RST-DT annotations for explicitly marked coherence relations that occurred more than 30 times in total (to keep the table readable). Mappings that were predicted by UniDim are presented in bold-font and underlined. The colors represent the values: greater numbers are shaded a darker red. Several clusters correspond well to the expected mapping; for example, PDTB and RST-DT labels that were decomposed as `temporal` or `causal` labels in UniDim map onto each other well. For the research goal of the current chapter, the discrepancies are more interesting. First, we can see that RST-DT's Circumstance was rarely annotated by PDTB as Con-

| | TEMP. | | CONT. | | COMP. | | EXPANSION | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **RST-DT label** / **PDTB label** | Synch. | Asynch. | Cause | Condition | Contrast | Concession | Conjunction | Instantiation | List | **Total** |
| Temp.-same-time | **63** | 1 | 1 | 1 | 4 | | 5 | | | 75 |
| Temporal-after | | **48** | | | 1 | | 2 | | | 54 |
| Sequence | 2 | **29** | 1 | | 1 | | 19 | | | 52 |
| Circumstance | **89** | 79 | 29 | 21 | 7 | 4 | **10** | | | 259 |
| Result | 1 | 3 | **30** | | 2 | | 5 | | | 41 |
| Consequence | 5 | 6 | **45** | | 4 | 1 | 16 | | 1 | 95 |
| Explanation-arg. | 6 | | **39** | | 7 | 2 | 1 | | | 57 |
| Reason | | 2 | **67** | | | | | | | 72 |
| Condition | 5 | 13 | 1 | **104** | 2 | 1 | 1 | | | 182 |
| Contrast | 1 | 1 | | | **160** | 23 | 17 | | | 208 |
| Concession | 5 | 4 | | 3 | 101 | **53** | 4 | | | 182 |
| Antithesis | 1 | 3 | | | **170** | **37** | 10 | | | 243 |
| Comparison | 2 | 1 | | | 26 | 2 | **9** | | | 41 |
| Elaboration-add. | 4 | 3 | 1 | | 30 | 8 | **122** | 3 | 3 | 192 |
| Example | 1 | | | | 1 | | 3 | **29** | | 35 |
| List | 2 | 4 | 1 | | 17 | 1 | 303 | | **47** | 377 |
| **Total** | 187 | 197 | 215 | 129 | 533 | 132 | 527 | 32 | 51 | |

**Table 4.4:** Alignment of explicit coherence relation classes for which $n$ >30. Numbers indicate how many instances occurred in the mapping; values are represented as colors. Underlined, bold numbers indicate the mapping as predicted by UniDim. Total numbers correspond to the total number of those labels in the data (including labels occurring less than 30 times).

junction. By contrast, these relations received an Asynchronous label in PDTB rather often. This does not conform to the description in the manual, which states that the events described are “somewhat co-temporal” (Carlson & Marcu, 2001, p. 48). The annotations of Circumstance therefore seem to deviate in practice from the manual’s definition.

It should be noted that RST-DT’s Circumstance does not have a directly corresponding label in PDTB. The same holds for RST-DT’s Comparison. Although there is no directly corresponding label, the manual does state clearly that the spans are not in contrast (Carlson & Marcu, 2001, p. 50). Based on this definition, the label was not mapped to PDTB’s Contrast in UniDim, but rather to Conjunction (a `positive additive` label). However, the mapping shows that Comparison is in fact often annotated as PDTB’s Contrast, and not as Conjunction.

There are also discrepancies in annotations for labels that do have a corresponding

| | TEM. | CON. | COM. | EXPANSION | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| RST-DT label / PDTB label | Asynch. | Cause | Contrast | Conjunction | Alternative | Instantiation | List | EntRel | Total |
| Background | **7** | 12 | 10 | 16 | 2 | 4 | | 21 | 72 |
| Circumstance | **1** | 10 | 8 | **11** | | 6 | | 15 | 51 |
| Consequence | 4 | **16** | 7 | 11 | 2 | 2 | 1 | 4 | 49 |
| Evidence | | **12** | 2 | 12 | 29 | 28 | | 6 | 92 |
| Explanation-arg. | 2 | **114** | 16 | 16 | 35 | 51 | 1 | 22 | 267 |
| Reason | 1 | **20** | 1 | 3 | 1 | 2 | | 3 | 31 |
| Evaluation | | **7** | 9 | 11 | 2 | **1** | | 12 | 46 |
| Interpretation | | **16** | 3 | 9 | 2 | 8 | | 8 | 49 |
| Contrast | 1 | 2 | **35** | 2 | | | 2 | 3 | 56 |
| Comparison | | 1 | 24 | 14 | | | 2 | 2 | 44 |
| Elaboration-add. | 29 | 168 | 73 | **221** | 36 | 151 | 7 | 266 | 991 |
| Example | | 9 | 1 | 6 | **64** | 21 | | 2 | 106 |
| Elab.-gen.-spec. | | 8 | | 11 | 17 | **44** | | 18 | 99 |
| List | 6 | 24 | 29 | 120 | 6 | 13 | **74** | 30 | 311 |
| Comment | | 16 | 9 | **8** | | **4** | | 8 | 49 |
| **Total** | 75 | 499 | 277 | 511 | 202 | 406 | 88 | 463 | |

**Table 4.5:** Alignment of implicit coherence relation classes for which $n$ >30. Numbers indicate how many instances occurred in the mapping.

counterpart in the other framework. First, consider RST-DT's Concession, which is often annotated as PDTB's Contrast. This result is similar to the result found in the PDTB–CCR comparison (Section 4.1), and can be attributed to difficulty in annotating this subtle distinction. Second, we can see that RST-DT's List relations are often annotated as PDTB's Conjunction. This difference stems from the different qualifications for what counts as a List relation: in PDTB, Lists have to be "announced"; in RST-DT, no such criterion is applied. The mapping of List relations is another example of how differences in operationalization can account for discrepancies between frameworks. Next, we turn to the mapping of implicit relations.

**Implicit relations.** We can infer from Table 4.5 that the agreement between frameworks is a lot lower on implicit relations than on explicit relations: the explicit relation mapping showed a diagonal line of shaded cells that largely overlap with the expected correspondences based on UniDim, but for the implicit relation annotation, a substantial proportion of relations from almost all PDTB classes were annotated as Elaboration-additional in RST-DT. This observation will be further discussed in the general discussion (Section 4.3). We now turn to a discussion of other clusters of discrepancies that Table 4.5 shows.

First, Table 4.5 shows that instances labeled as RST-DT's Evidence and Explanation-argumentative fall into PDTB's Instantiation and Restatement.Specification classes relatively often (approximately 40% of all Evidence and Explanation-argumentative relations). This finding is similar to what we found in the PDTB–CCR mapping.

Second, we see that there seems to be some confusion between RST-DT's Example and Elaboration-general-specific as well (the equivalents of PDTB's Instantiation and Specification): 20% of Examples are annotated as PDTB's Restatement.Specification, and 17% of Elaboration-general-specific are annotated as PDTB's Instantiation. These findings indicate that Instantiation and Specification relations can be quite ambiguous, and license further investigation. We will return to this issue in the discussion.

Finally, we again see that RST-DT's List relations are not consistently annotated as List in PDTB: Conjunction, Contrast and EntRel also occur frequently. Note that when PDTB annotated a List relation, RST-DT is often in agreement.

### Discussion PDTB–RST-DT mapping

The PDTB–RST-DT mapping led to several observations regarding how compatible different corpora and annotations are in practice, and how well UniDim can predict correspondences between annotations. Regarding the first type of observations, we find that differences in segmentation rules lead to difficulty in aligning the data. 24% of PDTB relations could not be mapped with certainty to a corresponding RST-DT label. This result highlights that segmentation has a strong effect on determining the scope and argument structure of a coherence relation (see also Hoek et al., 2017a). Although this finding is very relevant to the goal of the current study (mapping existing annotations of overlapping corpora), it is less relevant to the ultimate goal of UniDim: the alignment procedure in the current study showed that discourse structures and segmentation results differ between frameworks, but this does not mean that a researcher interested in Result relations cannot make use of items annotated with corresponding labels in both the PDTB and RST-DT. The alignment issues that were identified in this study therefore do not indicate major issues for making corpora more interoperable.

A second point relating to the compatibility of different corpora concerns annotations of explicit vs. implicit relations. Studies that report on inter-annotator agreement in coherence relation annotation efforts often do not distinguish between agreement on explicit and implicit relations (e.g., Carlson et al., 2003; Prasad et al., 2008). However, the current study revealed that agreement on these two types of relations differs greatly, at least between frameworks: while agreement on explicit relations is relatively acceptable, agreement on implicit relations is remarkably low. A similar result likely occurs for agreement within frameworks. For explicit relations, annotators can make use of the information that the relatively strong cues (i.e., connectives) provide; however, for implicit relations, annotators have to rely on less well-known and obvious signals, and the resulting annotations are hence based more on the annotators' subjective interpretations. In order to get more insight into this issue of difference in agreement, we recommend that future annotation efforts report agreement on explicit and implicit relations separately (as was also recommended in Hoek & Scholman, 2017).

Regarding the correspondences between UniDim's predictions and the annotations in practice, the results are fairly similar to those of the previous study: most mismatches did not occur due to wrong decomposition of the labels, but due to different operationalizations within the frameworks (e.g, of LIST relations). The results also revealed an additional cause for mismatches, namely omission: when a relational label is defined in one framework, but not in the other. Examples of such labels include RST-DT's BACKGROUND and CIRCUMSTANCE. Such labels are more unique than the "core" relation types (such as cause-consequence relations). A more detailed description of these labels could improve the mapping of these labels. However, it is possible that these RST-DT labels have a function that is not present in the PDTB scheme at all, and therefore such relations are not really annotated in the PDTB and cannot be mapped reliably.

## 4.3 General discussion and conclusion

The studies described in this chapter aimed at validating UniDim's proposed mapping on actual annotations. The results indicate that the expected correspondences match the observed patterns relatively well. Mismatches occur mainly due to (i) segmentation differences that lead to different interpretations of relations; (ii) differences in operationalization that lead to systematic disagreements (i.e., annotating a relation as `additive` INSTANTIATION or `causal` EXPLANATION-ARGUMENTATIVE; or annotating a LIST relation or a CONJUNCTION relation); and (iii) subtle, difficult to annotate distinctions between labels (e.g., CONTRAST vs. CONCESSION). On the basis of the analysis of the mapped data, we conclude that UniDim is a promising interlingua for unifying different discourse frameworks, and that discourse-annotated corpora have the potential to be interoperable.

**Disagreement due to operationalizations** The mapping studies have highlighted the difficulty of reaching agreement on a single label (between frameworks, but this also holds for efforts using a single framework, see Spooren & Degand, 2010; Hoek & Scholman, 2017). Disagreement is often considered to be the result of an error on the account of one of the annotators. However, in the case of agreement between frameworks, the mappings have shown that disagreement is, for a large part, due to the frameworks' operationalizations: the instructions often bias an annotator towards a particular label in order to facilitate the annotation process and increase the inter-annotator agreement. For future efforts, it is advisable to report explicitly about the annotation process and operationalizations, and how they might influence the resulting annotations.

The systematic discrepancies that are caused by differences in operationalizations can be attributed to theoretical differences between approaches. For the development of a general theory of coherence, such discrepancies warrant more investigation, in order to justify the approach taken. The next chapter (Chapter 5) discusses how the distinctions that are proposed in approaches can be justified. Specifically, it is argued that discrepancies can be evaluated by considering their cognitive plausibility; that is, whether the distinctions reflect the psychological constructs that are present in comprehenders' mental representation of texts.

**Disagreement due to ambiguities** A second cause for disagreement between annotators stems from the ambiguity of relations. This was also highlighted by Rohde et al. (2015, 2016): they found that, often, an additional sense can be inferred for relations containing an adverbial. They then suggest that some of the disagreement in coherence relation annotation might in fact be due to the possibility of inferring multiple readings for certain relations. When annotator A interprets a relation as INSTANTIATION, and annotator B interprets that same relation as CAUSE, both annotations might be correct, even though traditional agreement measures imply that one of these labels is "wrong" (see also Hoek & Scholman, 2017). Since the annotated data that are currently available do not carry multiple annotations, the community does not know how much of the data can be interpreted in different ways.[5]

Based on these findings, it seems that traditional annotation methods – every coherence relation is annotated by one or two expert coders – do not suffice, because they do not properly reflect the ambiguity of coherence relations. Moreover, one can wonder how reflective the annotations are of the naïve readers' interpretations, which might be different from an expert's interpretation. (Chapter 5 will reflect on the importance of having annotations reflect naïve readers' interpretations.) These issues imply that another methodology is needed that can supplement the existing annotation methods and open up new possibilities for discourse researchers. Ideally, this new method should be able to correspond to and supplement different frameworks, and at the same time reflect the naïve readers' interpretations. A suitable place to start for recruiting naïve participants and asking them to use such a method is on crowdsourcing platforms. Chapter 6 therefore introduces a new method for eliciting coherence relation interpretations from naïve, crowdsourced participants.

The results from the mapping studies that were discussed in this chapter function as input for the experiments in the remainder of the thesis. The disagreement regarding PDTB INSTANTIATION and SPECIFICATION relations is addressed in Chapters

[5] Approximately 5% of the data in PDTB 2.0 is double-annotated, but this likely does not reflect the real distribution of multi-functional relations.

7 and 8, using the crowdsourced experimental design that is described in Chapter 6. The finding that agreement on implicit relations is relatively low (compared to the agreement on explicit relations) motivates the studies in Chapters 9 and 10. Specifically, these studies aim to provide insight into whether comprehenders take into account both the connective and the content of the segments in order to infer coherence relations, or whether they rely more on the connective (if present).

Chapter 5

# Justifying distinctions between coherence relations

In the previous chapters, we discussed the proliferation of coherence relation frameworks and the relational categories that are distinguished in different frameworks. In this chapter, we will show that these approaches often focus on the descriptive adequacy of their inventories; that is, the inventory is developed to describe all relations in texts. Full-coverage theories of coherence relations aim to account for different types of observations; not only the intuitions from experts, but also insights from acquisition, production, and comprehension studies. For such theories, descriptive adequacy is not the only relevant measure; cognitive plausibility is also important. Theories that address both measures can improve our understanding of the mental processes of discourse representations, production, and comprehension. Descriptively adequate approaches can function as a starting point for developing such a general theory: they can provide an inventory of all possible relational constructs, which can then be validated (verified or falsified) using cognitive evidence.

This chapter makes the criterion of cognitive plausibility more tangible by detailing which sources of evidence can verify or falsify distinctions between coherence relational labels and classes. The cognitive status of features can be explored in a systematic and conclusive manner using this approach. The criterion of cognitive plausibility forms the motivation of the studies in the remainder of this thesis: specific distinctions that frameworks disagree on are evaluated by looking at whether naïve readers can make these distinctions, and whether it affects their interpretation processes.

## 5.1 Justification of coherence relation theories

Approaches that have proposed inventories of coherence relations generally agree on the coarse-grained relation types, but they differ in many important aspects regarding the types of fine-grained relations that are distinguished. It is often argued that choosing a single best set of relations that is then used by the entire community is not feasible, given that different research goals require different inventories (e.g., Taboada & Mann, 2006; Versley & Gastel, 2013). However, the development of a general theory of coherence relations would benefit the community: a full-coverage theory that can account for discourse structure in general can provide more insight into the constructs that are present in our linguistic system, and improve our understanding of the mental processes of discourse comprehension and production. Hence, even though research projects might make use of their own set of relations, a more general set would have a major impact on the field of discourse.

What distinctions should be part of such a theory? We consider existing frameworks to be a starting point for forming a full-coverage theory. Comprehensive theories are subject to stringent requirements and criteria, and to rigorous testing by others. By evaluating distinctions made in frameworks similar to how theories would be evaluated, we can improve our understanding of the phenomenon of discourse coherence. Evaluating and validating frameworks as theories should therefore not be viewed as criticism of a framework, but as an effort to contribute to a more general theory of coherence.

Theories that aim to explain language as it is represented in the mind can be held accountable for the hypotheses that they generate. The accountability relates to two principles: validity and reliability. These principles are most often discussed in relation to methodology, but they can also be applied to theories. A valid theory is soundly reasoned and logical. To determine the validity of a theory, one can ask: how accurately does the theory represent the construct? Testing the validity of a theory is related to falsification: researchers should not look for evidence that supports or confirms the theory, but rather look for evidence that the theory is not valid. This means that the hypotheses that can be developed based on theories should be falsifiable; that is, it should be possible to refute these hypotheses. Reliability is a measure of stability or consistency: can experts agree on a description of a linguistic object (such as an annotation of a relation) using the constructs of the theory? Another method is to test a theory's internal reliability: does an expert agree with his or her own earlier annotations using the constructs of the theory?

The discussion in the current chapter mainly relates to the validity of theories: in order to validate theories, the distinctions that they propose need to be justified. This can be done by trying to falsify the distinctions. The next sections elaborate on the process of justification. Two general measures of justification have been posited for linguistic theories: *descriptive adequacy* and *cognitive plausibility.* Each measure will be introduced, providing information on what the requirement means in general, and what it means for coherence theories in particular. We then focus on making the measure of cognitive plausibility more concrete.

### 5.1.1 Descriptive adequacy

**Origins** Descriptive adequacy is a quality measure for the evaluation of linguistic theories. It was first described by Chomsky (1965) as part of his theory of levels of adequacy in relation to grammar. He distinguishes three levels of adequacy for evaluating models: theories achieve *observational adequacy* if they account for a finite corpus of data; theories that have *descriptive adequacy* account for the linguistic intuitions of native speakers; and theories that achieve *explanatory adequacy* can explain how such knowledge is acquired and have predictive power.

Although Chomsky's theory of adequacy was originally proposed to account for the adequacy of grammars, it can be applied to discourse as well: it requires theories to account for the structure of a particular corpus of data, as well as the intuitions of native speakers. Brewer & Lichtenstein (1982) discuss the descriptive adequacy of several theories regarding the discourse structure of narrative stories. They tested this by asking participants to rate the naturalness of stories that adhered to one of the theories. If a certain story was considered as natural by the native speakers, they took this as evidence for the descriptive adequacy of the corresponding theory. Brewer & Lichtenstein (1982) therefore adhered closely to Chomsky's definition of descriptive adequacy by focusing on the intuitions of native speakers. Other discourse researchers have adapted the definition of descriptive adequacy by placing less of an emphasis on the intuitions of native speakers (considering that the experts that create the theories are also using their intuitions). For example, Sanders et al. (1992) consider theories to be descriptively adequate if they make it possible to describe the structure of all kinds of natural text. Similarly, Wolf & Gibson (2005) consider visualizations of discourse structures descriptively adequate when they can account for all different kinds of structures that an annotation study revealed.

**Descriptively adequate theories of coherence relations** Descriptively adequate coherence relational categories function as useful tools to describe text structure (Sanders et al., 1993). They are descriptive constructs, developed by linguists based on the texts that are available. Applying this criterion to theories of coherence relations in particular, one could argue that the theory must be able to describe all relations in the texts that it aims to analyze (Knott, 1996), or in other words, it should account for a particular corpus of data. This can be referred to as *observational completeness* (cf. Nuyts, 1992). For example, if a theory aims to describe the coherence relations in spoken discourse or in a specific genre, the relational inventory should be sufficient to enable a complete analysis of these texts. Additionally, the categories that the theory distinguishes should be distinguishable; i.e., a model should not distinguish categories of which no examples can be found in natural text. This can be referred to as *observational adequacy* (cf. Nuyts, 1992).

**Evidence for descriptive adequacy of a theory** As stated in the definition of a descriptively adequate theory of coherence relations, the theory must adhere to observational completeness: it should be able to describe all relations in the texts that it aims to describe. A full analysis of texts by using the theory could therefore be considered as evidence for the descriptive adequacy of the theory. Underlying this type of evidence is the agreement between multiple experts on whether the distinctions that the theory proposes do in fact correspond to the actual constructs in language. This is not to say that every linguist should agree that the proposed relational categories exist, but at least more than one expert should be able to distinguish the categories from each other. This agreement must exist in order for the theory to adhere to the intuitions of native speakers.

### 5.1.2 Cognitive plausibility

**Origins** Cognitive plausibility is related to cognitivism, which is the study in psychology that focuses on mental processes and seeks to understand cognition. A cognitivist view on language covers a number of dimensions, including a description of the adult cognitive systems and a description of the acquisition and development of the cognitive system in children (Nuyts, 1992).

In cognitive science, the term 'cognitive plausibility' has many different meanings. For example, in cognitive modeling, cognitive plausibility refers to being able to replicate the observed behavior of individuals, and in AI, the term refers to the inputs and outputs of systems being comparable to those of humans (Kennedy, 2009). In language acquisition, the usefulness of a computational model is considered to be directly related to how closely it approximates the actual acquisition task (Phillips & Pearl, 2015). This can be referred to as the cognitive plausibility of the model. What these fields have in common is that their models need to replicate human performance in order to be cognitively plausible.

**Cognitively plausible theories of coherence relations** Coherence relations are considered to be entities that comprehenders infer when they create a mental representation of the discourse; in other words, these relations are psychological constructs that are present in the comprehender's representation. Applying the measure of cognitive plausibility to theories of coherence relations means that the theory must be able to describe discourse as it is encoded in human cognition; that is, it should distinguish relational categories that actually play a role in the construction of a cognitive representation of text (Sanders et al., 1992).

**Evidence for cognitive plausibility of a theory** There are various sources of evidence that can provide justification for the existence of relational categories. For example, evidence can come from empirical studies (e.g., in the areas of acquisition, comprehension, production) that show that the proposed relational categories have different effects on comprehenders' cognitive processes. Specific examples of such studies are age-of-acquisition studies that show that additive connectives are acquired before causal connectives (Bloom, Lahey, Hood, Lifter & Fiess, 1980), or eyetracking-while-reading studies showing that objective causal relations are processed faster than subjective causal relations (Traxler, Sanford, Aked & Moxey, 1997). In Section 5.2, the different types of evidence that can be used to justify cognitive plausibility will be discussed in more detail, but first, we discuss the relation between descriptive adequacy and cognitive plausibility.

### 5.1.3 Descriptively adequate *and* cognitively plausible?

Descriptive adequacy and cognitive plausibility are not contradictory; rather, they involve two different dimensions of the same object of investigation (Nuyts, 1992). Theories and the relational categories that they distinguish can therefore be both descriptively adequate and cognitively plausible; researchers need not make a choice. In practice, however, coherence relation approaches often focus on one more than the other. In this section, we will discuss the three approaches that are most relevant to the work reported in this dissertation. For each approach, we will consider what their goal is (do they aim for descriptive adequacy in their inventory, or cognitive plausibility), what the object of investigation is (connectives, relational labels, or relational categories), whether they can be verified and falsified, and whether that is actually done.

Regarding the object of investigation, we make an explicit distinction between *relation labels* and *relational categories*. Relational labels are fine-grained labels that are proposed in certain approaches, such as in RST-DT. Relational categories are coarse-grained relation types that are proposed, such as `causal` and `additive` in CCR. In this section, we will discuss how these objects of investigation relate to PDTB, RST-DT, and CCR. The sources of evidence that are nominated later in this chapter will also be discussed according to whether they can be used to justify distinctions between labels or distinctions between categories, or both.

**PDTB** The PDTB is a theory-neutral approach that aims to be descriptively adequate. In the creation of PDTB's inventory, the authors took a data-driven, bottom-up approach to coherence relations: rather than prescribing relational labels and trying to "fit" the data in this prescriptive system, they took the lexicon of connectives as a starting point and identified relational labels that correspond to these connectives. The resulting inventory consists of relational labels that are organized in four more general relational categories, or *classes* (see Chapter 3 for a more detailed description of the relational inventory).

The relational inventory can be verified by considering the connective lexicon: the existence of a specific connective indicates the existence of a corresponding relation label. For example, the connective *except* motivates the distinction of a specific EXCEPTION relation. If multiple connectives can express a single relational label, this could suggest that the label can be further refined. For example, CONJUNCTIONS can be expressed by a variety of connectives in the PDTB 2.0, including *moreover* and *similarly.* In PDTB 3.0, a separate relational label (SIMILARITY) is distinguished for relations that can be expressed by *similarly.*

Relational distinctions could be falsified by using the connective lexicon as well. For example, if an approach were to distinguish a relation such as BANANA PEEL, whereby one segment introduces a banana peel in the discourse, and another expresses that somebody slipped on that peel, this distinction could be falsified by the fact that there is no connective in the lexicon that can uniquely express the BANANA PEEL relation (see also Knott, 1996).

However, it should be noted that the connective lexicon is in fact a source of evidence regarding the cognitive plausibility of the distinction, as we will see in Section 5.2 (see also Knott & Dale, 1994; Knott & Sanders, 1998). Using the connective lexicon to verify the descriptive adequacy of the relational labels therefore speaks towards the cognitive plausibility of these distinctions, more than to their descriptive adequacy. The PDTB does not propose other sources to verify or falsify distinctions.

**RST-DT** The RST-DT focuses on the descriptive adequacy of relational labels. The RST website emphasizes this focus on descriptive adequacy by stating that “RST is intended to describe texts, rather than the processes of creating or reading and understanding them.”[1]

RST-DT proposes relational labels to describe texts. These labels are grouped in 16 general classes, but these classes are not applied in annotation; that is, they only use the classes for organizing and presenting their relational inventory. RST-DT does not annotate connectives.

Justification of the relational labels in RST-DT is provided by their descriptive adequacy, in the sense that the RST-DT annotators used the labels to analyze texts and reached sufficient inter-annotator agreement. No further verification or falsification of the distinctions is provided or proposed.

Interestingly, even though RST-DT explicitly focuses on descriptive adequacy and not on cognitive plausibility, their relational distinctions do relate to cognitive insights: relation labels are defined as functional constructs associated with the particular effects that a writer intends to achieve. But no justification is provided that the relational labels that they distinguish are in fact cognitively relevant.

**CCR** Sanders, Spooren & Noordman (1992, 1993) developed CCR as a more general theory of coherence relations, aiming for cognitive plausibility. They explicitly did not aim at providing a “complete descriptively adequate taxonomy of coherence relations”

[1] `http://www.sfu.ca/rst/01intro/intro.html`

(Sanders et al., 1992, p.4). Instead, they suggest that researchers can create a more descriptively adequate inventory by further specifying the classes using additional features (such as those proposed in the UniDim approach, see Chapter 3).

CCR distinguishes relational classes using dimensions such as polarity (`positive` and `negative` relations) and *basic operation* (`causal` and `additive` relations). It is possible to combine the values for these dimensions in order to create a taxonomy of relational labels, but such a taxonomy is not used for annotation. Connectives are annotated as relevant linguistic signals of specific relational classes (e.g., *because* is a signal of causal relations).

The relational classes can be verified and falsified using different sources of evidence. Sanders et al. (1992) and Sanders et al. (1993) present a series of experiments that support CCR's distinctions. These experiments investigate the agreement between experts and interpretation by naïve participants. Moreover, they consider evidence from other studies looking at, for example, the age of acquisition of connectives and processing differences between relational types, as support for these dimensions.

The distinctions can be falsified by considering possible negative results from these experiments. For example, Sanders et al. (1993) used a sorting task (Miller, 1969) to investigate whether comprehenders are able to distinguish different types of relations from each other. They presented participants with pairs of discourse segments and asked them to sort these pairs on the basis of the similarity of the coherence relations. They hypothesized that participants would be able to distinguish between relations that differed in their values for the dimensions distinguished in CCR. The results showed that participants were indeed able to create clusters of `positive causal` relations, `positive additive` relations, `negative causal` relations, `negative additive` relations, and `conditional` relations. However, participants did not create separate clusters for `objective` and `subjective` relations, which speaks against the cognitive plausibility of this specific distinction. Such negative results indicate that the specific distinction warrants more investigation. The issue of falsification will be further discussed in Section 5.2.4.

It can be expected that approaches focus more on one measure than the other, depending on what is relevant to their research goals. For example, a theory of discourse as a linguistic object will aim to describe the structure of the text, without taking into account the cognitive status of this structure. Such a theory can be useful for a variety of tasks, such as automatic text summarization. The theory's cognitive plausibility is irrelevant for such a task, since text summarization usually does not relate directly to human processing tasks. The main focus is on extracting the most important information from the text. As a result, the theory can be good for that particular goal, even though it does not account for the cognitive plausibility of the distinctions that it makes.

When using descriptive adequacy as the only criterion for developing or justifying a more general, full-coverage theory, it can be difficult to assess the appropriate level of detail (see also Knott, 1996). For example, the PDTB distinguishes six subtypes of Condition relations (differing in truth status and tense, for example), and two more subtypes of Pragmatic condition relations. By contrast, RST-DT distinguishes four types of Condition relations. Both theories seem feasible as a descriptive framework, and both have been used to annotate texts, thereby conforming to the requirement of observational completeness. Descriptive adequacy does not provide any other external measures to verify or falsify these distinctions. Additional sources of evidence are thus needed to provide justification. These sources speak to the cognitive plausibility of the distinctions.

Another issue with using descriptive adequacy as the sole requirement for theories concerns its explanatory power. Descriptively adequate approaches are useful tools to *describe* discourse structure, but they are not meant to *explain* discourse structure. In other words, they do not have sufficient explanatory power. Theories with explanatory power make predictions about the object of investigation, and these predictions can be falsified. Descriptively adequate theories are not developed for hypothesis generation, and they cannot be falsified easily because of this. Cognitively plausible theories do allow for hypothesis testing, which makes them falsifiable.

For a full-coverage theory of coherence relations, descriptive adequacy and cognitive plausibility are therefore both important: only theories that address both measures can improve our understanding of the mental processes of discourse comprehension and production. Descriptively adequate approaches can function as a starting point for developing a general theory: they can provide an inventory of all possible relational constructs that comprehenders can produce and that experts can identify and agree on. Researchers can then validate (i.e. verify or falsify) descriptively adequate relational categories or labels using cognitive evidence. Cognitively plausible evidence for relational distinctions strengthens theories and provides justification that the theory models the actual constructs that are present in our linguistic system. Moreover, theories that make a stronger commitment to cognitive plausibility have more explanatory power, which allows researchers to make empirically testable predictions that can be verified or falsified. Testing such predictions can advance the field of discourse coherence.

In the next section, we aim to make the measure of cognitive plausibility more tangible by outlining possible sources of evidence that can verify or falsify the cognitive plausibility of distinctions between labels or classes.

## 5.2 Criteria for cognitive plausibility

There is no source that details what the criteria are for deciding whether a particular relational distinction or coherence framework is cognitively plausible. The literature generally states that theories have to be responsive to empirical results from cognitive research, thereby also providing support for the framework's structure and assumptions. However, the specific types of empirical results or cognitive research are not elaborated on. Moreover, it is not clear what counts as *enough* evidence for the cognitive plausibility of a distinction, or how distinctions can be falsified. In this section, we list possible sources of evidence and discuss some issues related to the criterion of cognitive plausibility.

### 5.2.1 Possible sources of evidence

The possible sources of evidence that are nominated are: production, (off-line) representation and comprehension, (online) processing, acquisition, linguistic evidence, annotation, and cross-linguistic evidence. These sources operate on different levels: some give more direct insight into the plausibility of relational distinctions (such as processing studies), others provide indirect insight into distinctions by studying connectives (such as linguistic evidence or age of acquisition studies). Some of the sources in this list prescribe methodologies (e.g., production, representation), others build on these sources by using the same methodologies but looking at different age groups (acquisition) or languages (cross-linguistic evidence). Moreover, some sources allow for hypothesis testing in order to try to falsify the distinctions (e.g., processing studies), whereas others can indicate possible issues with distinctions, but are less suitable for falsifying or verifying hypotheses (e.g., annotation studies).

The list proposed here is not exhaustive; it is merely meant to give an indication of the types of evidence that can be provided (see also Gilquin & Gries, 2009, for an overview of linguistic research methodologies). The methods that are discussed are exemplary for the methods that are commonly used in coherence relation research, and an example is given for most methods to illustrate the phenomena that can be investigated using the method. These function as pointers for those who are interested; discussing every example in detail is outside of the scope of this chapter.

**Production** Production data consists of written and spoken discourse. Descriptive adequacy is therefore related to this type of data: categories are distinguished based on the identification of corresponding relational constructs in written or spoken texts.

Production data mostly consists of produced texts, which are investigated in corpus-based studies. Corpora can contain different genres (e.g., newspaper, novels, broadcast interviews) and modalities (e.g., spoken, written, chat). Researchers can use this data to test hypotheses based on distributional data. For example, Andersson & Spenader (2014) investigate the occurrences of Result and Purpose relations that appear with and without their typical connective *so* in written corpora. The results show that Purpose relations always occur with *so*, whereas Result relations are often left unmarked. This distributional variability indicates that there is a significant difference between the two types of relations. The data is therefore taken to support the distinction between Purpose and Result relations.

Another source of production evidence comes from elicited production studies such as story continuation or connective insertion experiments. In story continuation experiments, participants are presented with prompts and asked to continue the prompt in a natural manner. This type of study is useful to test whether readers are more likely to infer a specific type of relation depending on different contexts. For example, Simner & Pickering (2005) investigate the planning of causes and consequences in language production using the story continuation paradigm; specifically, how do comprehenders decide which section of the prior discourse to target for continuation and how do they decide on the causality content of that continuation? The results show that comprehenders tend to anchor upcoming utterances to information that is both temporally recent and textually recent, and they seek to satisfy gaps in their discourse model by providing a cause when the prior discourse provided a consequence and vice versa. This indicates that comprehenders do not simply have uniform preferences for producing causal relations; rather, they select between cause and consequence based on fine-grained features of the discourse.

In connective insertion experiments, participants are presented with the segments of a relation and asked to insert a connective. Sanders et al. (1992) use such a task to investigate whether readers are able to infer the coherence relations between sentences and to express them by the appropriate connectives. They presented readers with a set of implicit relations that were originally explicit, and with a set of connectives to choose from. The results showed that participants were able to choose the connective corresponding to the original connective relatively well. Discrepancies between the participants' choices and the original connectives could be described in terms of the theory: a discrepancy often concerned a choice for a related class (such as the class of `negatives` for *but*) instead of an unrelated class (such as the class of `positives` for *but*). Most disagreements between the inserted connective and original connectives occurred for subjective relations. This indicates that the distinction between objective and subjective relations warrants more investigation.

**Representation and comprehension (off-line)** Certain off-line measures provide insight into the representation of the discourse after the comprehension process is completed. Methodologies tapping into these representations include recall, judgment tasks (e.g., acceptability or plausibility ratings) and comprehension questions. For example, Murray (1997) investigated whether adversative connectives contribute differently to the discourse than causal and additive connectives by asking participants to rate on a five-point scale whether sentences "followed sensibly" from the preceding sentence. The results showed that sentences with incorrectly placed adversative connectives such as *however* led to greater disruption in the perceived coherence than incorrectly placed additive and causal connectives.

Another type of experiment that can provide insight into the interpretation of relational types is the sorting task discussed in Section 5.1.3 (Sanders et al., 1993, see also Miller, 1969), for which participants sorted relations based on their similarity to each other. Such a task can provide insight into how comprehenders interpret relation types.

**Processing (online)** Various paradigms can be used to capture the online comprehension processes. Self-paced reading, eyetracking-while-reading, and the visual world paradigm are often used for such research goals. For example, Keenan, Baillet & Brown (1984) used a self-paced reading design to investigate the processing time of two-sentence items with different degrees of causal relatedness (e.g., "*Joey's big brother punched him again and again. The next day his body was covered with bruises.*" versus "*Joey went to his neighbor's house to play. The next day his body was covered with bruises.*"). The results showed that reading times for the second sentences steadily increased as the causal relatedness of the two sentences decreased, indicating that causally linked information is easier to process.

Wei (2018) used the visual world paradigm to investigate whether comprehenders are sensitive to the subjectivity profile of connectives. Participants listened to causal relations marked by a subjective or objective connective while looking at two images: one image depicted an event, the other depicted a person with a speech bubble containing a picture of the event. The results showed that participants looked at the image of the person more often after hearing a subjective connective than an objective connective. This indicates that the subjectivity of a connective has an immediate influence on readers' expectations of the relation. The results of this study therefore support the cognitive plausibility of the distinction between objective and subjective relations.

Measuring reading times or proportion of looks are a rather natural way of measuring comprehension: the interpretation process is not disturbed by the experimental task. However, reading times can sometimes be ambiguous regarding the types of processes they index; for example, longer reading times do not tell us what type of difficulty causes it. Neurological measures such as ERP methods provide a clearer view of what processes occur during reading. For example, Drenhaus, Demberg, Köhne & Delogu (2014) complemented results from a visual world study conducted by Köhne & Demberg (2013). In this visual world study, Köhne & Demberg (2013) investigate the processing of causal and concessive connectives. They find that readers are able to quickly process and integrate causal connectives in the discourse, but not concessive connectives. The authors hypothesize that these results may be due to an increased difficulty of processing concessives, or of integrating them with upcoming content. Drenhaus et al. (2014) followed up on this experiment with an ERP study. The data show that, when readers encounter a concessive connective instead of a causal connective, they update their mental representation from an expected causal relation to an unexpected concessive relation (as reflected by a higher P600 for processing a concessive connective). This specific updating process supports the explanation of the visual world study that their results are due to an increased difficulty of processing concession compared to causality.

**Acquisition** Evidence from acquisition studies is rather perpendicular to the sources discussed so far: the same methodologies are used to investigate a different group of language users. Children acquire connectives and coherence relations in a relatively fixed order (e.g., Bloom et al., 1980). The system that underlies this order is expected to correspond to a cognitively plausible categorization of relations (Sanders et al., 1992). Many studies that have investigated the acquisition of discourse coherence have focused on the acquisition and comprehension of connectives (rather than relations), because an understanding of the meaning and use of connectives indicates that the child has also acquired the meaning and use of the type of relation that corresponds to the connective.

A number of different paradigms have been used to study child language acquisition. Perhaps the simplest method for studying the acquisition of discourse skills is by recording children's spontaneous speech (Ambridge & Rowland, 2013). Such recordings can be found in corpora of longitudinal child language data, such as the CHILDES corpus (MacWhinney, 2014). These corpora have provided a fruitful source of production evidence. For example, Evers-Vermeul & Sanders (2009) built on Bloom et al's (1980) work by investigating the order of acquisition of connectives in the CHILDES corpus. Looking at Dutch data, they found that additive connectives such as Dutch *en* ('and') are acquired earlier than causal connectives as well as adversative connectives (such as 'but'). These results provide evidence for the cumulative cognitive complexity theory, which states that additive relations are less complex than causal relations, which in turn are less complex than adversative relations.

Production methods such as elicited-production experiments also provide researchers with a method to investigate discourse coherence skills in children. There are many different types of tasks in this paradigm, and they differ in how structured they are. Generally, however, the task consists of asking a child to respond to some sort of question or stimulus. For example, Evers-Vermeul & Sanders (2011) use a directive task to investigate different domains of causality in children's speech: children were asked to instruct a puppet where to put stickers and to motivate their instruction. The results showed that children aged two were able to use causal connectives in the objective domain, but that the subjective domain is acquired later.[2]

Another source of acquisition evidence comes from comprehension and processing methods, such as act-out, picture-pointing or visual-world studies. For example, Knoepke et al. (2017) investigated children's comprehension of causal and adversative (or negative causal) sentences using a verification task: children aged 6 to 10 years old and adults were presented with auditory stimuli and asked to judge the coherence of the sentence pairs. The results showed that both children and adults were less accurate in judging the coherence of adversative relations compared to causal relations. The study provides further evidence for the cumulative cognitive complexity theory: causal relations are less complex than adversative relations.

[2] Evers-Vermeul & Sanders (2011) used the term *content* for the objective domain and *epistemic* and *speech act* for the subjective domain.

**Linguistic evidence** Relations can be expressed or marked in different ways, leading to different sources of linguistic evidence. The most commonly considered and explicit source is cue phrases: the existence of a cue phrase in a language indicates the existence of a corresponding relational category. Knott & Dale (1994) created a framework of coherence relations based on the existence of cue phrases indicating the proposed categories. They used substitutability tests – which tap into people's intuitions about whether one phrase can replace another in a given context – to investigate the connections between different cue phrases and their corresponding relational categories (see also Knott & Sanders, 1998).

The existence of cue phrases can also be used to validate other theories. For example, if a certain framework distinguishes several subtypes of a relation, which are all typically marked by the same cue phrase (e.g., different types of Condition relations all marked by *if*), this could be considered evidence that the subtypes might not be cognitively plausible. Similarly, if a given relational category is typically marked by a variety of cue phrases, the category might be too coarse-grained and could be further distinguished into subcategories.

Cue phrases and connectives are generally considered the strongest markers of coherence relations, and they have been at the center of research in the signaling literature. However, the encoding of a relation can happen on different levels; not only on the connective level (see Stukker & Sanders, 2012). It is possible that the connective lexicon does not provide evidence for a specific type of relation, but that such relations are in fact encoded differently (such as in the syntactic features). The most comprehensive collection of signals is the RST-DT Signalling Corpus (Das & Taboada, 2018), in which coherence relations have been annotated for all possible signals: reference, semantic, lexical, syntactic and graphical features (see also Hoek, 2018, Chapter 5). Future studies will hopefully provide more insight into the link between these other types of markers and specific relational categories.

**Annotation** Annotation processes and output could provide interesting (albeit indirect) insights into the plausibility of categorizations. After all, a theory of coherence relations must propose valid relational categories that can be used to analyze texts and can be distinguished reliably by annotators.

One source of annotation evidence is how easily items can be annotated according to this distinction. Taboada & Mann (2006) refer to this as *observability* (the possibility of distinguishing one relation from another). As an example, they discuss the Elaboration relations in the RST-DT. The theory distinguishes six different types of Elaborations, but it has been found to be relatively difficult to decide on which specific subtype holds in practice. A similar argument has been made regarding the distinction between the subtypes of PDTB's Contrast relations. Annotation practice has shown that it is very hard to distinguish reliably between Contrast.opposition and Contrast.juxtapostion, and so the updated version of PDTB (version 3.0) does not distinguish between these subtypes anymore (Webber et al., 2016). The annotation process (ease with which a distinction can be made) and outcome (amount of agreement between annotators) can therefore be used as evidence that specific categorizations are not entirely valid or reliable.

The mapping studies reported in Chapter 4 also provide a source of annotation evidence: the discrepancies that were found in the mapped data indicates that the categorizations of those relations might not be optimal. For example, PDTB's INSTANTIATION and SPECIFICATION relations were often annotated in RST-DT as EVIDENCE and EXPLANATION-ARGUMENTATIVE. Although the mapping does not indicate which categorization is more plausible, it does signal that these distinctions warrant more investigation. These results can therefore be used as input for further studies. Annotation evidence can thus be considered to be indirect evidence for the cognitive plausibility of a theory.

**Cross-linguistic evidence** Finally, studies investigating (the realization of) discourse coherence phenomena in multiple languages can also provide valuable insight into the plausibility of a category. If a distinction is found to be cognitively relevant in one language but not another, this could indicate that the distinction needs more investigation.

First, however, we should discuss a more fundamental question: are relational categories language-specific? There is no consensus on an answer yet. Although more general classes of relations are argued to hold across languages (e.g., CAUSAL, CONTRASTIVE), more fine-grained relational distinctions that are made in theories of coherence seem to differ between languages. Consider the PDTB, which has been adapted to different languages such as Arabic (Al-Saif & Markert, 2010), Italian (Tonelli, Riccardi, Prasad & Joshi, 2010), and Chinese (Zhou & Xue, 2015). These efforts have all made changes to the relational inventory proposed by the PDTB. For example, Zhou & Xue (2015) merge certain relational categories for their Chinese framework (e.g., LIST and CONJUNCTION were merged because the distinctions were often hard to make during the annotation process) and PROGRESSION and PURPOSE were added.

The adaptations of frameworks to different languages can have two reasons. First, it is possible that fine-grained relational distinctions that were identified for one language might not be relevant in all other languages. This would mean that relational categories can be language specific, in which case more research is needed to fully comprehend which relational distinctions are affected, and why they are found to be relevant in one language but not another. It could be that a particular aspect of a culture creates the need to express a certain type of relation that is not relevant in other cultures. Second, it is possible that the differences are caused by differing expert intuitions (regardless of language), rather than actual linguistic differences. In this case, relational categories are not language-specific, and cross-linguistic evidence can be used to identify which distinctions warrant more investigation. It is of course also possible that both explanations hold. Future empirical studies investigating the plausibility of distinctions in different languages might provide more insight into this issue.

Cross-linguistic evidence can come from studies using the methodologies discussed so far (i.e., the same methodologies can be used to investigate different languages). Another valuable source of cross-linguistic production evidence is translation studies. There tends to be a lot of variation in the way coherence relations are expressed in the target language compared to the original source language (Halverson, 2004; Zufferey & Cartoni, 2014). Researchers can therefore determine whether the marking of coherence relations varies cross-linguistically by considering multiple translation pairs. For example, Hoek et al. (2017b) investigated which factors make coherence relations more, or less, likely to remain implicit by conducting a parallel corpus study. They hypothesized that cognitively simple relations are more often implicit than relations that are cognitively more complex, based on the Cognitive approach to Coherence Relations. The results showed that cognitive complexity indeed influences the linguistic marking of coherence relations, and that this does not vary between the languages in their corpus.

Cross-linguistic differences between coherence relational distinctions is a promising area of future research. Most studies to date have looked at the meanings of discourse connectives cross-linguistically (Zufferey & Degand, 2013), translations of connectives to different languages (e.g., Hoek & Zufferey, 2015), and relations between connective lexicons for different languages (e.g., Bourgonje, Grishina & Stede, 2017; Knott & Sanders, 1998). These studies have provided interesting insights into the different lexicalizations of relations in various languages. Less is known about different relational distinctions and the effects of such distributions on processing and acquisition in different languages (but see Drenhaus et al., 2014; Köhne & Demberg, 2013; Mak, Tribushinina & Andreiushina, 2013; Sun & Zhang, 2018; Zufferey, Mak & Sanders, 2015, among others). Looking at cross-linguistic differences in the types of relations that exist, how they are marked and distributed, and how they are processed is important for general theories of discourse coherence, since they ideally should be able to account for more than one language.

In sum, several sources of evidence are nominated for justifying relational distinctions based on the criterion of cognitive plausibility: evidence from production, comprehension, and processing studies, as well as from the (cross-)linguistic system, annotation efforts and the field of acquisition can provide support to the cognitive possibility of relational categories. Of course, these different sources do not provide the same type of insights. The observability that is taken into account during annotation efforts provides indirect insight into the plausibility of categorizations, as does evidence from lexicons and cross-linguistic studies. Other sources give more direct insight into the plausibility of categories; consider various production and comprehension methodologies. The different insights that can be gained using these methods are exactly what makes the combination of evidence stemming from various methodologies informative regarding the plausibility of distinctions.

**Falsification** The process of falsifying distinctions deserves some more attention. Some methodologies that were proposed allow for hypothesis testing to falsify the distinctions. Others are less suitable for direct testing; they can indicate that a distinction is not valid, but they cannot directly verify or falsify.

Corpus-based studies can be used to develop hypotheses, as well as to try to falsify them. Hypotheses can be developed by considering relational distributions; for example, by looking at the type of relations that negation words can mark (Asr & Demberg, 2015). Hypotheses can also be tested by considering relational distributions (see, for example, Asr & Demberg, 2013).

Linguistic evidence can be used for forming hypotheses (e.g., if there is a connective that prototypically marks a specific relation type, that relational category is cognitively plausible) and for testing hypotheses (e.g., if a relational category is cognitively relevant, there must be a connective that can prototypically mark that category) (see Knott & Dale, 1994; Knott & Sanders, 1998).

Elicited production studies, processing experiments, and representation and comprehension studies are mainly useful for testing hypotheses, as done by Sanders et al. (1992): the hypothesis that comprehenders can infer the subjectivity of implicit coherence relations was not supported by the connective insertion experiment. Acquisition studies are also useful for testing hypotheses. Cross-linguistic studies might be useful for falsifying distinctions (e.g., if a distinction is cognitively plausible in one language, it should also be cognitively plausible in other languages), but it should first be investigated how feasible it is to require fine-grained relational categories to be cognitively relevant across all languages.

### 5.2.2 Applying the criteria as a case study

In order to illustrate how distinctions can be evaluated using the more defined criterion of cognitive plausibility, Table 5.1 presents an overview of the evidence that has been found to support (or contradict) the cognitive plausibility of the Unifying Dimensions proposal; a more detailed version of the table (including references to the relevant sources) can be found in Appendix E.[3] In the remainder of this section, we first make general observations regarding the data in Table 5.1 (see also Appendix E) and then focus the discussion one specific feature (*Alternative*).

First, the table shows that more studies have investigated the plausibility of the first five rows, which represent the dimensions, compared to features. This does not necessarily mean that these features are not cognitively plausible; it only means that more research is needed in order to determine their cognitive status. The difference

[3] This overview is by no means complete; for example, many more studies can be cited as evidence that causal and additive relations are cognitively plausible distinctions. However, it can be assumed that for those cells that remain empty (NA), the literature study revealed no relevant sources.

| **Dimension / feature** | **Evidence from** | | | | | | |
|---|---|---|---|---|---|---|---|
| | Production | Repres./Compr. | Processing | Acquisition | Linguistic system | Annotation | Cross-linguistic |
| Polarity | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Basic operation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Source of coherence | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ |
| Implication order | ✓ | NA | ✓ | ✓, ✗ | ✓ | ✓ | ✓ |
| Temporality | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Specificity | NA | NA | NA | NA | ✓ | ✓, ✗ | ✓ |
| List | NA | NA | ✓ | NA | ✓ | ✗ | ✗ |
| Alternative | NA | NA | NA | NA | ✓ | ✓, ✗ | ✓ |
| Goal-orientedness | NA | ✓ | NA | NA | ✓ | ✓ | ✓ |
| Conditionality | ✓ | NA | NA | ✓ | ✓ | ✓ | ✓ |

**Table 5.1:** An overview of the evidence from various sources regarding the cognitive plausibility of distinctions in the Unifying Dimensions proposal. ✓: evidence; ✗: counter-evidence; NA: not available. See Appendix E for more details regarding the sources.

in the number of studies can be attributed to their respective grain size: dimensions apply to more general classes of relations, whereas features apply to fine-grained distinctions within these more general classes. The focus of studies has mainly been on the general classes, in order to show that coherence is cognitively plausible. Given the amount of evidence available for these general classes, it is our hope that the field will move towards studying more fine-grained distinctions in the future.

Second, regarding the contributions from various sources, we conclude from Table 5.1 that most research investigating these dimensions and features has studied the linguistic system. Representation studies exploring distinctions between relational classes or labels are least frequently conducted. Moreover, it should be noted that evidence from annotation studies was not always available because many efforts do not report agreement on finer-grained distinctions or confusion matrices; instead, they report an average agreement score. The evidence that is included in Table 5.1 mainly comes from the mapping studies reported in Chapter 4. We hope that, in the future, distinctions will be justified with experimental data from these less common sources as well, in order to further investigate the cognitive plausibility of the labels and classes.

We now focus our discussion mainly on one feature in order to keep the discussion brief. The *Alternative* feature is chosen because both evidence and counter-evidence have been found, and some methodologies have not yet been used to investigate the plausibility. The *Alternative* feature helps to distinguish additive relations in which the two segments are presented as alternatives from additive relations in which this is not the case. PDTB 2.0 distinguishes several relations within the group of *Alternative* relations; specifically, DISJUNCTIVE, CONJUNCTIVE and CHOSEN-ALTERNATIVE. DISJUNCTIVE relations are exclusive alternatives (*either...or...*), CONJUNCTIVE relations are inclusive alternatives (*and/or*); these relations thus differ in their polarity. RST-DT groups DISJUNCTIVE and CONJUNCTIVE relations together under the label DISJUNCTION, and it does not distinguish CHOSEN-ALTERNATIVE.

As Table 5.1 shows, evidence for the group of *Alternative* relations (and the distinctions made within) does not come from all sources: while several studies have identified or investigated unique linguistic markers for these relations (such as *instead* or negation words; see, e.g., Asr & Demberg, 2015; Knott, 1996; Webber, 2013), a literature review revealed that no work has been done on the acquisition, representation, comprehension, and processing of disjunction on the discourse level compared to other types of relations (but see Staub & Clifton, 2006). Moreover, many annotation efforts have not explicitly reported agreement on *Alternative* relations and the subclasses within. The PDTB–CCR annotation effort in Chapter 4 forms an exception: agreement on DISJUNCTION relations was relatively high between the PDTB 3.0 and CCR annotations, which supports the distinction of DISJUNCTION relations. It should be noted, however, that in PDTB 3.0, the subclasses DISJUNCTIVE and CONJUNCTIVE are merged, which speaks against the observability of this specific subclassification.

In sum, the strongest evidence for the cognitive plausibility of the ALTERNATIVE feature comes from the linguistic marking of these relations. However, evidence from other sources is lacking and future studies will hopefully address these gaps in order to provide more support for the justification of *Alternative* relations. For example, age of acquisition studies can provide insight into whether the connectives marking these relations occur at a unique and typical stage during development. Processing studies can provide insight into whether DISJUNCTIVE relations require more processing effort than CONTRAST relations. Both types of relations are `additive` and `negative`, but CONTRAST relations do not present alternatives. Also, sorting tasks can be used to study if comprehenders consider DISJUNCTIVE, CONJUNCTIVE and CHOSEN ALTERNATIVE to be different from each other.

This discussion of evidence for the *Alternative* feature and, more generally, the overview shown in Table 5.1 raise several issues that deserve consideration from the community and will be elaborated in the next sections. First, given that many features do not have supporting evidence from all sources, one can wonder what counts as enough evidence? Second, we can see that in some cases, sources have provided counter-evidence for certain distinctions (e.g., low observability regarding the *Source of coherence*). This raises the question of what counts as “enough” counter-evidence to conclude that a feature is not cognitively plausible?

### 5.2.3 Verification: what counts as enough evidence?

Ideally, the distinctions made by theories are justified by available evidence from all sources. However, it is likely that a theory cannot account for all behavioral phenomena within its range, because we simply do not know everything related to cognition. To address this, Nuyts (1992) suggests to speak of the *probability* of a theory being cognitively plausible: the more data that a theory accounts for, the higher the probability that it is an adequate characterization of discourse coherence. From this point of view, any evidence of a distinction being cognitively plausible is a good start, but there really is no way to say that there is enough evidence. Perhaps "sufficient" is a better term.

Support for a particular distinction should come from more than one source, and should ideally accumulate over time. Evidence from only a single study can be considered indicative of the existence of the category, and will need to be supplemented with evidence from other sources, using different methodologies, participants and materials. This idea is consistent with earlier proposals for providing converging evidence from multiple methods before considering a theory to be valid. Hobbs (2003) already noted that, ultimately, a psychological theory will have to be responsible for all different types of data. Magliano & Graesser (1991) and Graesser, Singer & Trabasso (1994) advocate a "three-pronged method" that coordinates (i) predictions generated by theories or hypotheses, (ii) data from think aloud protocols (a paradigm in which subjects are asked to verbalize what they think as they read or write), and (iii) behavioral measures that asses processing time. In a similar vein (albeit recommending different methodologies), Sanders & Evers-Vermeul (to appear) advocate a combination of (i) corpus studies on language use, (ii) experimental studies on discourse processing and representation, and (iii) corpus-based and experimental studies on language acquisition. They argue for this combination of converging methodologies because each methodology has its own merits and drawbacks, and they can therefore complement each other.

### 5.2.4 Falsification: what counts as evidence that a distinction is not valid?

As noted previously, researchers should aim to falsify distinctions; that is, to find evidence that the distinction doesn't exist. The question that arises is: when is a distinction falsified? One answer would be to adhere to a checklist: all sources of evidence discussed in Section 5.2.1 should be investigated. A lack of evidence in studies using one of these methodologies then indicates that the distinction isn't valid. In other words, if one were to commit fully to the criterion of cognitive plausibility, an ideal framework would be one that only includes relational categories for which empirical, cognitive evidence has been found in all measures.

However, this is a dangerous approach. Cognitive discourse research has not been conducted for very long yet, and it is likely that future studies will uncover more cognitive aspects of coherence relations. If studies have failed to produce evidence supporting a particular type of relation, this could be due to many different factors. Of course, it could be that this particular category does not exist. But it could also be that the category might not affect cognitive processing as much, or that we haven't found a measure or method that is sensitive enough to pick up on any effect. We should not be too quick to throw out distinctions that are made based on descriptive adequacy. Basing a measure on only cognitive plausibility and excluding descriptive adequacy is therefore not an approach that we would recommend.

Nevertheless, we need a more defined idea of what counts as evidence, and what counts as counter-evidence. In the previous subsection, we stated that evidence from one source can be considered indicative of the existence of the distinction. Continuing this line of reasoning, we argue that evidence from one source indicating that the distinction does not affect cognition should be considered as counter-evidence for the cognitive plausibility of the distinction. However, this should then be verified by evidence from multiple sources, using different methodologies and materials. Even then, we cannot be certain that the distinction is not cognitively plausible, we should only consider the probability of the plausibility to be low.

A more problematic issue concerns those distinctions for which both evidence and counter-evidence can be found. Consider, for example, the class of adversative relations, and in particular the distinction between CONTRAST (i.e., semantic opposition) and CONCESSION (i.e., negative causal) relations. Theories have provided different ways to distinguish these relations from each other (Izutsu, 2008; Prasad et al., 2007; Carlson et al., 2003; Pander Maat, 1998), and they have provided linguistic examples as justification for their categorizations. However, agreement on the classification of relations in these two types of categories is very hard to reach (Zufferey & Degand, 2013, see also Demberg et al., submitted). This would indicate that the distinction is not a functional one. Looking at other sources of evidence, though, the distinction does seem to be relevant. For example, there is a general consensus that both categories have typical connectives: *whereas* and *by contrast* are typical contrastive connectives, and *even though* and *nevertheless* are typical concessive connectives (Knott, 1996; Prasad et al., 2007)

There is no easy solution for what to do with such distinctions. Given the amount of evidence in favor, one could conclude that the probability of this distinction is being cognitively plausible is relatively high. Perhaps the issue is with its descriptive adequacy: the distinction between CONTRAST and CONCESSION might not be optimally categorized yet. For example, Robaldo & Miltsakaki (2014) argue that previous classifications do not account for the full range of variants of CONCESSION relations, and they propose a different categorization based on the source of expectation (e.g., *correlation* and *implicature* are two different sources). They then report a significant improvement on the annotation of previous disagreements on CONCESSION-CONTRAST relations in the PDTB. This indicates that the new classification might be a more plausible one, although this of course warrants more investigation. Distinctions for which there's conflicting evidence might require more investigation to determine whether the classification is the most optimal one. This is the type of debate that the community would benefit from, and that could advance the field of coherence.

## 5.3 Conclusion

In this chapter, we have discussed different methods of providing justification for theories of discourse coherence, and various sources of evidence for cognitive plausibility in particular. We have tried to make the criterion of cognitive plausibility more defined, and hope that this will contribute to future research efforts aimed at justifying relational distinctions. Moreover, we hope that the issues raised in this chapter will spark a debate about the evaluation of distinctions between coherence relation types.

Both measures – descriptive adequacy and cognitive plausibility – are based on the notion that distinctions made in theories should correspond to and account for linguistic data. But there are many different types of data that can be used to evaluate proposed distinctions between coherence relation types: human language is produced, comprehended, and acquired, which can be investigated by analyzing (monolingual or translation) corpus data or by conducting production, representation, processing or acquisition experiments. We here argue that providing converging evidence stemming from all of these sources is crucial for verifying the validity of theories of coherence relations. Corpus data can provide insight into how language is produced, and thus allows experts to identify constructs that converge with descriptive adequacy. Proposed constructs can be verified with experimental data, to ensure that they are cognitively plausible. Crucially, we do not argue that descriptive adequacy is not a necessary requirement; rather, a text-analytic model that aims to have explanatory power should meet both requirements (or rather, meet descriptive adequacy and strive for cognitive plausibility).

There are still several points that are open to discussion. First, how should we view distinctions with conflicting (i.e., evidence and counter-evidence) evidence? The proposal in the current chapter is to reconsider their classification: distinctions such as Contrast vs. Concession might require further investigation to determine whether the current classification is the most optimal one. The question is whether this solution works for all distinctions with evidence and counter-evidence. Second, the issue of cross-linguistic evidence deserves more consideration in the community: is it feasible and valid to expect relation types to be (equally) cognitively plausible in all languages? Or should we expect differences in the types of fine-grained relation types that can be distinguished in various languages?

On a more general level, the community would benefit from a discussion of the grain size of relations that might be supported by evidence from these sources. Specifically, we can expect a discrepancy between the relation types that experts and naïve language users can distinguish. How should we deal with such variability? Should we exclude fine-grained relation types that naíve language users cannot distinguish from a full-coverage theory of coherence relations? Or should we make a distinction between relation types that only experts can identify and those that most language users can identify?

In sum, we have argued for a greater role for cognitive plausibility in the justification and evaluation of coherence relation distinctions that are made in theories and frameworks. Moreover, we have defined an approach for this justification by detailing what the criterion of cognitive plausibility entails. Such an approach is relevant for the next chapters, where discrepancies in annotations are investigated by looking at cognitive data. Specifically, the pattern of disagreement between PDTB and RST-DT annotators regarding INSTANTIATION and SPECIFICATION relations found in Chapter 4 is evaluated by investigating whether readers interpret the relations that are identified by PDTB and RST-DT annotators.

In Chapters 9 and 10, we then turn to a different issue in theories of coherence: whether comprehenders can build and maintain an expected discourse structure over multiple sentences. This is investigated cross-linguistically, in order to evaluate whether any effects are modulated by language-specific factors. Hence, assumptions that are made in theories regarding the structure of relations are evaluated by considering cognitive evidence.

## Chapter 6

# Investigating the influence of context and the reliability of a connective insertion task[1]

Traditional discourse annotation tasks are often costly and time-consuming, and their reliability and validity are in question. In this chapter, we develop and evaluate a new crowdsourcing method for which participants are asked to choose a connective from a predefined list that can express the connection between the segments of a relation. Relations were taken from the PDTB corpus to facilitate the evaluation of the data. This method allows us to elicit naïve readers' discourse interpretations, from which discourse annotations can be deduced. A second research goal is to investigate the effect of linguistic context on the reliability of the task, in order to provide more insight into the optimal design of such a task.

The results showed that the majority of the inserted connectives expressed the same meaning as the original PDTB label encoded. The distribution of connectives revealed that more than one sense can often be inferred for a single relation. The results showed no significant effect of context on the connective insertions. However, a by-item comparison revealed segment characteristics that influenced the annotations. This knowledge can be used in the design of coherence relation annotation tasks. The findings discussed in this chapter indicate that the newly developed crowdsourcing method has the potential to function as a reliable alternative to traditional annotation methods. Moreover, the distributions of inserted connectives provide evidence that relations can have multiple senses, which has important implications for future discourse annotation efforts.

[1] This chapter is adapted from, and in some parts identical with, Scholman, M.C.J. and Demberg, V. (2017). Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. *Proceedings of the 11th Linguistic Annotation Workshop (LAW)* (pp. 24–33). Valencia, Spain.

## 6.1 Introduction

In order to make further progress in the study of coherence relations, researchers need large amounts of annotated data, and these data need to be reliable and valid. However, manually annotating coherence relations is a difficult task that is prone to individual variation (Spooren & Degand, 2010) and requires a large amount of time and resources. Because researchers try to find a balance between obtaining reliable data and sparing resources, many projects are characterized by the standard practice of using two trained, expert annotators to code data.

This procedure is time-consuming and costly, and it raises questions regarding the reliability and validity of the data. When using trained, expert annotators, they may agree not because they follow instructions, but because they share implicit knowledge and know the purpose of the research well (Artstein & Poesio, 2008; Riezler, 2014). By including more annotators who are not experts in the annotation process, researchers can better ensure the reliability of the data (Krippendorff, 2004)

In this chapter, the feasibility of crowdsourcing discourse annotations is investigated. Employing non-trained, non-expert (also referred to as naïve) subjects to obtain annotations allows us to collect large amounts of coded data in a short period of time. Moreover, the process ensures that the obtained annotations are independent and do not rely on implicit expert knowledge; instead, the task allows us to directly tap into naïve subjects' interpretations.

It should be noted, however, that crowdsourcing has rarely been used for coherence relation annotation efforts. This could be attributed to the nature of crowdsourcing: typically, crowdsourced tasks are small and intuitive tasks, and so participants do not receive extensive training. Under these conditions, crowdsourced annotators cannot be asked to code according to a specific framework (this would require them to study manuals). Therefore, rather than asking participants to assign relation labels to instances, we ask them to choose a connective from a predefined list. In order to ensure that these connectives are as unambiguous as possible (see Asr & Demberg, 2013), we chose connectives based on a classification of connective substitutability by Knott & Dale (1994). We investigate how reliable the obtained annotations are by comparing them to expert annotations from two existing corpora.

A second research aim is to examine the effect of the design of the task on the reliability of the data. It is generally assumed that coherence relations should be supplied with their linguistic context in order to ensure reliable annotations. However, there are no clear guidelines for or investigations of how much context is needed for reliable annotation. The current study therefore examines the influence of context on readers' interpretations of coherence relations.

## 6.2 Background

Working with naïve annotators has several advantages: they are easier to come by, which makes it easier to employ a larger number of annotators. This, in turn, decreases a possible effect of annotator bias (Artstein & Poesio, 2005, 2008). Moreover, naïve annotators are not influenced by their experience with and knowledge of a specific framework, which ensures that a framework bias or instruction bias will not occur in the annotations. They are also not experts when it comes to discourse coherence, which ensures that their annotations will not rely on expert knowledge (Riezler, 2014).

Various studies have found high agreement between naïve and expert annotators for Natural Language tasks (e.g., Snow et al., 2008). However, annotating coherence relations is a different and especially difficult type of task, because they require complex semantic interpretations of relations, and because textual coherence does not reside in the verbal material, but rather in the readers' mental representation (Spooren & Degand, 2010). Despite this increased difficulty, naïve annotators have recently also been employed successfully in coherence relation annotation tasks conducted by Scholman et al. (2016) and Kawahara et al. (2014).

Scholman et al. (2016) developed a systematic, step-wise annotation scheme based on CCR to obtain annotations from naïve in-lab participants. These participants were given instructions containing connective substitution and relational paraphrase tests that were designed to guide them through the annotation process. The results showed that two of CCR's four dimensions, *polarity* and *order of the segments*, could be applied reliably by non-trained annotators. The other two dimensions, *basic operation* and *source of coherence*, were more problematic. The results also showed that the substitution and paraphrase tests led to higher agreement. Scholman et al. (2016) concluded that non-trained, non-expert annotators can be employed for discourse annotation, that a step-wise approach to coherence relations based on cognitively plausible principles is a promising method for annotating discourse, and that text-linguistic tests can guide annotators during the annotation process.

By using in-lab annotators, Scholman et al. (2016) were able to have annotators work with a small manual and instructions. However, crowdsourced annotators – unlike expert or naïve in-lab annotators – are not typically asked to code according to a specific framework because this would require them to study manuals, which goes against the small-scale characteristic of crowdsourced tasks. Kawahara et al. (2014) were able to employ crowdsourced annotators without having to train them. The authors developed a two-stage approach to obtain coherence relation annotations. First, participants were asked to determine whether a relation held between two segments; next, they were asked to label the relation by considering various examples of every label. The tagset consisted of causal and adversative labels. Kawahara et al. (2014) obtained 10 observations per relation and used an item-response model to determine which label was the "true" label for every item based on the observations (see Chapter 2 for an explanation of this evaluation approach). Using this method, Kawahara et al. (2014) were able to create a discourse-annotated corpus comprising 30,000 sentences in less than eight hours. This approach is interesting, but unlikely to be maintainable with more labels: asking participants to compare each item to examples of a large number of labels is inconvenient. It would eventually be the same as asking them to study a new framework.

In the current study, a different approach is taken. Instead of asking participants to assign specific relation labels to relations, we ask them to choose a connective from a predefined list of connective phrases. Of course, connectives are known to be able to mark multiple types of relations. Consider *but*, for example, which can mark CONTRAST, CONCESSION and ALTERNATIVE relations (see Asr & Demberg, 2013; Degand, 1998; Hovy, 1995; Maschler & Schiffrin, 2015; Versley, 2011, among many others). For the current study, we therefore aimed to choose connectives that are as unambiguous as possible, by relying on a classification of connective substitutability by Knott & Dale (1994), among other sources.

It is assumed that the connectives included in this experiment are prototypical markers of specific relational classes. However, we do not assume that, for example, it is impossible for an additive marker to also imply a causal reading. Instead, the current method rests on the assumption that participants will choose the connective that matches the strongest reading that they infer. This means that if they infer a causal reading for a specific relation annotated as CONJUNCTION by PDTB annotators, they will choose a causal connective. They can choose two connectives if they believe that both readings hold. Rohde et al. (2015, 2016) used a similar methodology to obtain insights into naïve participants' interpretations of relations (see also Sanders et al., 1992; Scholman et al., 2016).

Rohde et al. (2016) showed that readers can infer an additional reading for a coherence relation connected by an adverbial. By obtaining many observations for a single fragment, patterns of co-occurring relations could be identified; for example, they showed that readers often infer an additional causal reading for a relation marked by *otherwise*. These results highlight a problem that is characteristic of data annotated by two coders: differences in annotations might be written off as annotator error or disagreement, rather than reflective of the multiple interpretations for a relation (without there being a single correct interpretation). An approach that facilitates the use of many annotators is therefore more sensitive to the possibility of relations having multiple readings.

The current study uses a similar method as Rohde et al. (2016), but applies it to answer a different type of question. Rohde et al. (2016) investigated whether readers can infer an additional sense for a pair of sentences already marked by an adverbial. They did not have any expectations on whether there was a correct answer; rather, they set out to identify patterns of connective insertions. The current chapter explores whether crowdsourcing can be used to obtain annotations that are similar in quality to annotations provided by experts. The included items are taken from the Penn

Discourse Treebank. Crucially, we assume that there is a correct answer: the original label assigned by expert annotators. In order to evaluate the method's reliability, the connective insertions are compared to the original PDTB annotations.

It should be noted that inserting a connecting phrase is different from assigning a relation label. The annotations that we can obtain from connective insertions are more coarse-grained. However, the use of connectives for annotation does have several advantages of traditional annotations: the method can tap into the interpretations of relations by many different annotators, thereby reducing a possible effect of individual biases. Moreover, the method reveals a distribution of senses for a single relation, thereby reflecting the possible multiple interpretations of a relation (cf. Cuenca & Marín, 2009; Rohde et al., 2015, 2016; Webber et al., 2001).

**Context** The second research goal of the current study is to investigate how an annotation task for naïve subjects should be designed. Obtaining more insight regarding design decisions is crucial, given the limited amount of research into using naïve subjects for coherence relation annotation. One aspect of this design is the inclusion of context.

The benefits of context are well described in the field of discourse analysis. In particular, context is necessary for readers to ground the discourse that they construct (Cornish, 2009). The interpretation of every sentence (other than the first) in a text is therefore constrained by its preceding context (Song, 2010). It is argued that context has significant effects on important parts of discourse annotation; for example, determining the rhetorical role that each sentence plays in a discourse, and determining temporal relations between the events (Lascarides, Asher & Oberlander, 1992; Spooren & Degand, 2010). Because of these uses of context, it is assumed that the knowledge of context is a necessary requirement for discourse analysis.

Intuitively, it seems logical for context to be able to affect readers' interpretations of relations, especially for additive relations and in cases where the coherence relation at stake is also part of a relationship with that context. Additive relations such as CONJUNCTIONS are considered to have a weaker link between the segments, compared to causal relations (see, e.g., Sanders et al., 1992). For such additive relations, the context might therefore be more likely to steer the reader towards a different or stronger interpretation, compared to causal relations, which already have a strong link. Consider the following examples:

(18) The girls went to see Beyoncé in concert.
*Lilou cried.* **Annie beamed happily.**

(19) The two sisters were fighting again. Their mother took Annie's side.
*Lilou cried.* **Annie beamed happily.**

In Example 18, a CONJUNCTION relation holds between the segments in italics and bold. In Example 19, the segments are the same, but the relation itself is part of a different relation with the context: the segments both present a consequence of the mother taking Annie's side. The segments in italics and bold can therefore also be interpreted as a LIST relation that presents two consequences of an event in the context. Thus, even though we are interested in the relation between the adjacent segments, the relation of these segments with the context can have an effect on interpretations.

Few studies to date have investigated the role of context in discourse annotation. Sanders (1997) made an important contribution by exploring whether the presence of context influenced readers' annotations of the source of coherence of relations. In this study, participants were presented with relations whose source of coherence was either ambiguous, exclusively objective, or exclusively subjective. Participants were asked to indicate the source of coherence for every relation by choosing an appropriate paraphrase from a predefined list of paraphrases. The results showed that, for ambiguous relations, the context strongly determined the interpretation of that relation: the participants interpreted relations as objective when they appeared in descriptive contexts (that is, encyclopedia texts), and as subjective when they appeared in argumentative contexts. The type of context did not affect the interpretation of exclusively objective or subjective relations. These results suggest that comprehenders rely on context during discourse interpretation when the relation is ambiguous.

A second study investigating the influence of context on discourse interpretation was conducted by Canestrelli, Mak & Sanders (2016). In an eyetracking experiment, they investigated whether evaluative (subjective) markers such as *terribly*, *fantastic*, and *really* in the context facilitated the processing of subsequent subjective causal relations. Crucially, this study showed no effect of context. These results are not necessarily contradictory to those found by Sanders (1997), given that his study only revealed an effect of context for ambiguous relations.

The results of both of these studies indicate that subjectivity is mostly an aspect of the inherent relational meaning, which is not determined by the context. To date, we know very little about the influence of context on the interpretation of various other types of relations. The current study explores whether the presence of context generally has an effect on discourse interpretation (e.g., more agreement on a specific reading), without manipulating cues in the context.

Despite the presumed benefits of context, there are no clear guidelines for how much context is needed during discourse annotation. Consequently, studies have diverged in the methodology. In some projects, coders annotate the entire text (e.g., Rehbein et al. 2016; Zufferey et al. 2012). They will therefore likely take the context of the relation into account automatically during linear annotation. However, for projects where the entire text does not have to be annotated, or the task is split into smaller tasks, relations (or connectives) are usually annotated with a limited amount of context preceding and following the relation (e.g., Hoek & Zufferey, 2015; Scholman et al., 2016).

From a theoretical standpoint, the question of whether and how context influences readers' interpretations is interesting, because there are still many open questions about how readers construct coherence relations and mental representations of discourse. Practically, more insight into the effects of context on readers' interpretations is crucial because it will benefit coherence relation annotation efforts. Knowing how much context is minimally needed to be able to reliably annotate data will save resources; after all, the less context annotators have to read, the less time they need to spend on the task. The goal of the current experiment is therefore to test the agreement between crowdsourced discourse annotations and original corpus annotations, as well as the effect of context on the agreement between annotators.

## 6.3 Method

A crowdsourcing experiment was conducted for which naïve (non-trained, non-expert) annotators were asked to insert connectives from a predefined list into coherence relations taken from the PDTB. The items were divided into several batches. Each batch contained items with context or without context, but these two conditions were not mixed.

### 6.3.1 Participants

167 native English speakers (age range 20-68 years; mean age 36 years; 87 female) completed one or more batches of this experiment. They were recruited via Prolific and reimbursed for their participation (2 GBP per batch with context; 1.5 GBP per batch without context). Participants came from the United States, United Kingdom, Ireland, and Australia. Their education level ranged between an undergraduate degree and a doctorate degree.

### 6.3.2 Materials

The experimental items consisted of 234 relations from Wall Street Journal texts (see Appendix F). These relations were annotated by both PDTB (Prasad et al., 2008) and RST-DT (Carlson et al., 2003) annotators, and therefore carry both labels. The following types of PDTB relations were included: 24 CAUSE, 24 CONJUNCTION, 36 CONCESSION, 36 CONTRAST, 54 INSTANTIATION and 60 SPECIFICATION relations. Of the 234 relations, 192 were implicit and 42 were explicit. Only CONCESSION and some CONTRAST relations were explicit, since there were no implicit CONCESSION relations and too few implicit CONTRAST relations in the source texts carrying both PDTB and RST-DT annotations.

For all relation types besides INSTANTIATION and SPECIFICATION relations, the PDTB and RST-DT annotators agreed on the label. INSTANTIATION and SPECIFI-

[1] Crowdsourcing platform, `www.prolific.ac`

cation were chosen to accommodate the experiment reported in the next chapter, and for most of these, the PDTB and RST-DT annotators were not in agreement. We therefore also expect lower agreement on these relations in the current experiment.

The 234 items were divided into 12 batches, with 2 Cause, 2 Conjunction, 3 Concession, 3 Contrast, 4 or 5 Instantiation and 5 Specification items per batch. Order of presentation of the items per batch was randomized to prevent order effects. Subjects were allowed to complete more than one batch, but saw every item only once. Average completion time per batch was 16 minutes with context and 12 minutes without context. Due to presentation errors in one Conjunction, two Cause, and two Concession items, the final dataset for analysis consists of 229 items.

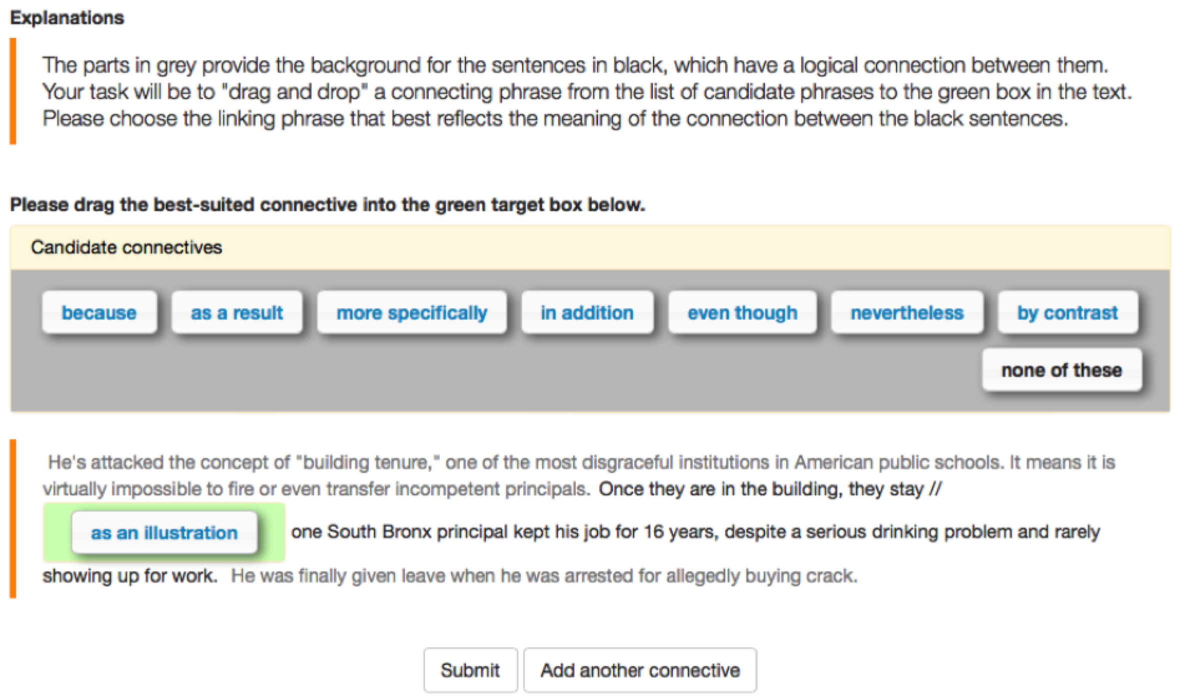
**Connecting phrases** Subjects were presented with a list of connectives and asked to insert the connective that best expresses the relation holding between the segments. The connectives were chosen to distinguish between different relation types as unambiguously as possible, based on an investigation of connective substitutability by Knott & Dale (1994). The list of connecting phrases consisted of: *because*, *as a result*, *in addition*, *even though*, *nevertheless*, *by contrast*, *as an illustration* and *more specifically*.

*Because* and *as a result* were chosen to represent Cause relations. *Because* is considered to be a general causal connective, underspecified for the source of coherence (in other words, it can be used to express objective as well as subjective relations). *As a result* is more debatable: Knott & Dale (1994) consider the phrase to be a prototypical marker of objective causal relations, whereas Cohen (1987) argues that this phrase can also be used to mark argumentative (subjective) relations. In the current study, we follow Cohen (1987). This will be reflected on in the Conclusion.

*Even though* and *nevertheless* were included as markers of Concession relations. *In addition* was chosen over *and* for Conjunctions because *and* is an underspecified connective and can be used to mark different types of relations, whereas *in addition* only marks additive relations. *By contrast* was chosen as a typical marker of Contrasts. Other typical markers, such as *but*, *although* and *however* were not included because they can be used to mark both concessive and contrastive relations.

Knott & Dale (1994) list *for example* as a typical marker of Instantiation relations. However, we decided to choose *as an illustration* instead. The two connecting phrases are interchangeable, but *as an illustration* is a slightly "heavier" connecting phrase, which might entice participants to make a more conscious decision when inserting it. Knott & Dale (1994) did not list markers for Specification relations. We chose *more specifically* as a typical marker.

Similar to the order of the items, the order in which the connecting phrases were presented was also randomized for every item.



**Explanations**

The parts in grey provide the background for the sentences in black, which have a logical connection between them. Your task will be to "drag and drop" a connecting phrase from the list of candidate phrases to the green box in the text. Please choose the linking phrase that best reflects the meaning of the connection between the black sentences.

**Please drag the best-suited connective into the green target box below.**

Candidate connectives

He's attacked the concept of "building tenure," one of the most disgraceful institutions in American public schools. It means it is virtually impossible to fire or even transfer incompetent principals. Once they are in the building, they stay // as an illustration one South Bronx principal kept his job for 16 years, despite a serious drinking problem and rarely showing up for work. He was finally given leave when he was arrested for allegedly buying crack.

**Figure 6.1:** Example of the interface of the experiment in the context condition. In this case, the participant chose *as an illustration* to indicate the relation between the segments.

## 6.3.3 Procedure

The experiment was distributed via Prolific and hosted on LingoTurk (Pusse, Sayeed & Demberg, 2016). First, participants were presented with instructions for the study. Next, they were presented with the experiment interface, which consisted of three parts: a short summary of the instructions, a box with predefined connectives, and the text passage (see Figure 6.1 for an example of the interface). In the context condition, the text passage contained two context sentences preceding the first segment and one context sentence following the second segment. These context sentences were taken from the original text and were not altered in any way. The two segments of the target relation were shown in black text while the context sentences were displayed in grey text. Subjects were instructed to choose the connecting phrase that best reflected the meaning between the black text elements, but to take the grey text into account. In the no-context condition, the grey sentences were not presented or mentioned.

Punctuation markers following the first argument of the relation were replaced by a double slash (//, cf. Rohde et al., 2015) to avoid participants from being influenced by the original punctuation markers (i.e., they might not insert the connective *because* due to a full stop after the first argument). The second argument always started with a lowercase letter.

In between the two arguments of the coherence relation was a box. Participants were instructed to “drag and drop” the connecting phrase that “best reflected the meaning of the connection between the arguments" (cf. Rohde et al., 2015) into this green box. Participants could also choose two connecting phrases if both phrases reflected the meaning of the relation, using the option "add another connective". Moreover, they could manually insert a connecting phrase by clicking "none of these" if they felt that none of the predefined options suited the relation. Participants were allowed to complete more than one batch, but they were never able to complete the same batch in both conditions.

## 6.4 Results

Prior to analysis, 5 participants from the context condition and 4 participants from the no-context condition were removed from the analysis because they had very short completion times (<10 minutes for 20 passages of 5 sentences each; <5 minutes for 20 passages of 2 sentences each) and showed high disagreement with other participants. The following analyses do not take the responses of these participants into consideration. In total, each list was completed by 12 to 14 participants. The total dataset considered for analysis consisted of 5886 observations.

As with any discourse annotation task, some variation in the distribution of insertions can be expected. We are therefore interested in larger shifts in the distribution of insertions. To evaluate these distributions, we report percentages of agreement (cf. De Kuthy et al., 2016). To evaluate the differences between conditions, we calculate Krippendorf's Alpha[2] ($\alpha$, Krippendorff, 1980), Fisher's exact test (Fisher, 1922) and Shannon's entropy (Shannon, 1948).[3]

We aggregated frequencies of the connectives that fell into the same class: *because* and *as a result* were aggregated as causal connectives, and *even though* and *nevertheless* were aggregated as concessive connectives.

In the next section, we first show evidence that the method is reliable. We then turn to the reliability of the no-context condition in comparison to the context condition to be able to determine whether the presence of context led to higher agreement on the sense(s) of items. Finally, we look at the entropy per item and per condition.

### 6.4.1 Overall reliability

The results indicate that the method is successful: the connectives inserted by the participants are consistent with the original annotation. This is shown in Figure 6.2a, with the bars reflecting the inserted connective per original class and condition. Figure 6.2b shows this distribution in more detail by displaying the percentage of

[2] Alpha was calculated using the R package agree.coeff2.r.

[3] Typically, annotation tasks are evaluated using Cohen's or Fleiss' Kappa (Cohen, 1960; Fleiss, 1971). However, Kappa is not suitable for the current task because it assumes that all coders annotate all fragments.

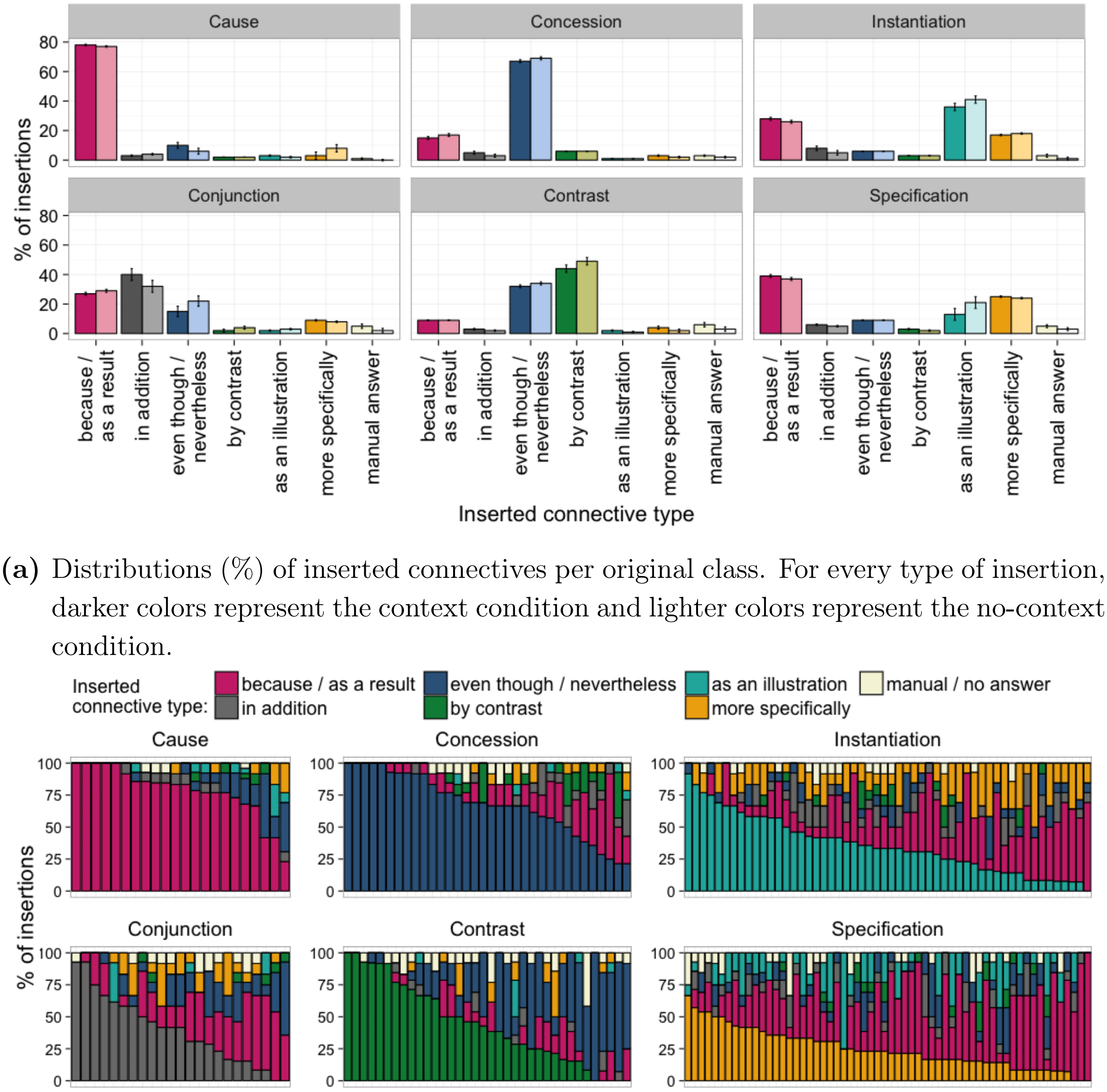
inserted connectives per item for the context condition. The distribution for the no-context condition is not included since it is almost identical to the distribution of the context condition. Every stacked bar on the x-axis represents an item; the colors on the bars represent the inserted connective.

These visualizations reveal several trends. First, for CAUSE and CONCESSION relations, the insertions often converge with the original label. 78% of the inserted connectives in items with a causal original label were causal connectives, and 67% of the inserted connectives in concessive items were concessive connectives. For both classes, the second most frequent category of inserted connectives was the other class: for CAUSE, the second most frequent category was CONCESSION (10%), and for CONCESSION, the second most frequent category was CAUSE (15%). On closer inspection of the items, we find that the disagreement between crowdsourced annotations and original annotations can be traced back to difficulties with specific items, and not to unreliability of the workers: the main cause for the confusion of causal and concessive relations can be attributed to the lack of context and/or background knowledge, especially for items with economic topics. For these topics, it can be very hard to judge whether a situation mentioned in one segment is a consequence of the other segment, or a denied expectation.

The second pattern that Figure 6.2a reveals concerns the classes CONJUNCTION and CONTRAST. The distribution of inserted connectives for these classes look similar: the expected marker is used most often (40% and 44%, respectively), with the corresponding causal relation as the second most frequent inserted connective type (27% causal insertions and 32% concessive insertions, respectively). A closer look at the annotations for items in these classes reveals that this is due to genuine ambiguity of the relation. For relations originally annotated as additive, we find that oftentimes a causal relation can also be inferred. The same explanation holds for CONTRAST relations: relations from this class that often receive concessive insertions are characterized by the reference to contrasting expectations. Some confusion between these relations is expected, as it is known that concessive and contrastive relations are relatively difficult to distinguish even for trained annotators (see, for example, Robaldo & Miltsakaki 2014; Zufferey & Degand 2013).

Finally, looking at INSTANTIATION and SPECIFICATION relations, we can see that there is more variety in terms of which connective participants inserted. This was expected, as these relations were chosen because the original PDTB and RST annotators did not agree on their annotation.

Looking at the no-context condition in Figure 6.2a, we find a near-perfect replication of the insertions in the context condition, with the distribution being stable on an item-by-item basis. This is further evidence for the reliability of the task. The agreement between the participants in the current study and the participants in the replication study is high: $\alpha = .71$. On average, the difference between the experiments on agreement with the PDTB label differed only by 3.7%. Fisher exact tests showed no significant difference in the distribution of responses between conditions



**(a)** Distributions (%) of inserted connectives per original class. For every type of insertion, darker colors represent the context condition and lighter colors represent the no-context condition.

**(b)** By-item distributions (%) for the context condition. Every bar represents a single item; the colors on the bars represent the inserted connective. Plots are arranged according to the number of dominant insertions corresponding to the original label.

**Figure 6.2:** Distributions (%) of inserted connectives per original class.

| **Original class** | **Context** | **No context** |
|---|---|---|
| Cause | 91 | 95 |
| Conjunction | 52 | 35 |
| Concession | 85 | 79 |
| Contrast | 53 | 58 |
| Instantiation | 54 | 46 |
| Specification | 25 | 20 |

**Table 6.1:** Percentage agreement between the original label and the dominant response per condition.

for any of the original classes (Cause: $p = .61$; Conjunction: $p = .62$; Concession: $p = .98$; Contrast: $p = .88$; Instantiation: $p = .93$; Specification: $p = .85$). Importantly, the results show that the distribution of insertions for every item is stable when two different crowdsourced groups take part in the experiment. Twelve insertions per item therefore seems to be an adequate amount to get a representative distribution of senses.

Another notable pattern, shown in Figure 6.2b, is that items often did not receive only one type of inserted connective; rather, they received multiple types of insertions. For Instantiation and Specification items, for example, participants often converged on two senses: both the originally annotated sense, as well as a causal reading. This indicates that multiple interpretations are possible for a single relation.

The data can also be analyzed by assigning to each relation the label corresponding to the connective that was inserted most frequently by our participants (in Figure 6.2b, this corresponds to the largest bar per item). This is known as the majority vote system (Hovy et al., 2013). We can then calculate agreement between the dominant response per item and the original label. These results are reported in Table 7.1.

Table 7.1 shows that the dominant response converges with the original label often for Cause and Concession relations and a majority of the time for Conjunction (in the context condition), Contrast and Instantiation relations (in the context condition). The dominant response for Specification items hardly converges with their original classification. This is as expected, as PDTB and RST-DT annotators also showed little agreement on Specification relations.

Looking at the effect of context, we see that agreement between the dominant response and the original label is slightly higher when context is present for four of six types of relations. For Conjunction relations, the agreement is even 17% higher in the context condition compared to the no-context condition. These results suggest that presence of context can have an influence on the subjects' interpretations of the relations. In the next sections, we will look at the distribution of individual items in more detail.

### 6.4.2 Effect of context: Dominant response per item

For 9% of the items, the dominant response shifts from one category to another depending on the presence of context. Manual inspection of these items revealed several characteristics that they have in common. First, it was found that, often, the topic is introduced in the context, and the (lack of) knowledge of the topic influenced the subject's interpretation of the relation when the context was absent. This is illustrated using the following CONJUNCTION example. In line with the PDTB, the first argument of the relation (Arg1) is indicated with italic font and the second argument (Arg2) with bold font. The additional text is the linguistic context.

(20) Quite the contrary – it results from years of work by members of the National Council on the Handicapped, all appointed by President Reagan. You depict the bill as something Democratic leaders "hoodwinked" the administration into endorsing.
*The opposite is true: It's the product of many meetings with administration officials, Senate staffers, advocates, and business and transportation officials* // **many congressmen are citing the compromise on the Americans With Disabilities Act of 1989 as a model for bipartisan deliberations.** Most National Council members are themselves disabled or are parents of children with disabilities. wsj_694

In Example (20), the context introduces the topic. The first argument then presents one argument for the claim that the bill results from years of hard work (as mentioned in the context), and the second argument is another argument for this claim. However, without the context, Arg2 can be taken as a result of Arg1. While this interpretation might be true, it does not seem to be the intended purpose of the relation. In the context condition, subjects interpreted the relation as a CONJUNCTION relation (58% of insertions were *in addition.* In the no-context condition, however, the dominant response was causal (58% of insertions), and the conjunctive *in addition* only accounted for 17% of all insertions.

Another common characteristic in items for which the presence or absence of context changes the dominant response, is that the context sentence following the relation expands on Arg2, thereby changing the probability distribution of that relation. This is common in INSTANTIATION and SPECIFICATION relations, where the second argument provides an example or specification of Arg1. Often, the sentence following Arg2 also provides an example or further specification, which emphasizes the INSTANTIATION/SPECIFICATION sense of the relation between Arg1 and Arg2. However, in relations for which Arg2 can also be seen as evidence for Arg1, the following context sentence can function to emphasize the causal sense of the relation by expanding on the argument in Arg2. Consider Example (21), taken from the class SPECIFICATION.

(21) Like Lebanon, and however unfairly, Israel is regarded by the Arab world as a colonial aberration. Its best hope of acceptance by its neighbours lies in

reaching a settlement with the Palestinians.
*Like Lebanon, Israel is being remade by demography* // **in Greater Israel more than half the children under six are Muslims.**
Within 25 years Jews will probably be the minority. wsj_1141

In this example, the context sentence following Arg2 expands on Arg2. Together, they convey the information that although Jews are the majority now, within 25 years Muslims will be the majority. Without the context, one could imagine that the text would go on to list more instances of how the demography is changing. Subjects in the no-context condition indeed seem to have interpreted it this way: 75% of the inserted connective phrases were *as an illustration*, and the remaining insertions were *even though* and *because*. By contrast, in the context condition subjects mainly interpreted a causal relation (64% of insertions), together with the specification sense (17%). The marker *as an illustration* only accounted for 7% of completions. Hence, with context present subjects interpreted Arg2 as providing evidence for Arg1, but without context it was interpreted as an INSTANTIATION relation.

### 6.4.3 Effect of context: Entropy per item

Another way of analyzing the influence of the presence of context on the participants' response, is to look at the entropy of the distribution of insertions. In the context of the current study, entropy is defined as a measure of the consistency of connective insertions. When the majority of insertions for a certain item are the same, the entropy will be low, but when a certain item receives many different types of insertions, the entropy will be high.

We calculated Shannon's entropy for every item. We then compared the conditions to determine whether entropy of an item increased or decreased depending on the presence of the context. Here we discuss items that have a difference of at least 1 bit of entropy between the conditions. This set consists of 18 items. Interestingly, presence of context only leads to lower entropy (higher agreement) in 10 items. For the other 8 items, subjects showed more agreement when the context was not presented.

*When context is beneficial* An analysis of items for which presence of context led to higher agreement revealed two common characteristics. First, similar to what we found in the previous section, presence of context is helpful when the context introduces important background information, or when the first argument refers to an entity or event in the context.

Second, we observed that agreement was higher in the context condition when Arg1 consists of a subordinate clause that attaches to another clause in the context. In these cases, the dependency of Arg1 to the context possibly hinders a correct interpretation of Arg1. Consider the following SPECIFICATION relation:

(22) The spun-off concern "clearly will be one of the dominant real estate development companies with a prime portfolio," he said. For the last year, Santa Fe

> Pacific has redirected its real estate operations toward longer-term development of its properties,
> *hurting profits that the parent had generated in the past from periodic sales from its portfolio* // **real estate operating income for the first nine months fell to $71.9 million from $143 million a year earlier.**
> In a statement late yesterday, Santa Fe Pacific's chairman, Robert D. Krebs, said that Santa Fe Pacific Realty would repay more than $500 million in debt owed to the parent before the planned spinoff. wsj_1330

In this example, Arg1 is a deranked subordinate clause, which cannot be used as an independent clause. All subjects in the context condition inserted a causal connective. However, in the no-context condition only 58% inserted a causal connective, and 33% of inserted connectives were *in addition*. Hence, the dominant response remained the same, but the amount of agreement decreased when the context was absent.

*When context is disadvantageous* Of the 8 items for which absence of context led to more agreement, 7 had a common characteristic: the relation between the context and Arg1 is not strong, for example because Arg1 is also the start of a new paragraph, or because there is a topic change. It is likely that in these cases, the presence of context took the focus away from the relation.

### 6.4.4 Double insertions and manual answers

Participants were given the option of inserting two connecting phrases if they thought that both phrases reflected the meaning of the relation. 3.4% of all answers consisted of two connecting phrases. For most items that received a double insertion, only one answer consisted of a double insertion. The data on multiple insertions therefore does not allow us to draw any strong conclusions. This will be elaborated on in the discussion.

When participants did not think any of the provided connecting phrases suited the relation, they were allowed to provide a manual answer. 2% of all insertions were manual answers. There was no clear pattern in these manual answers: only a few items received manual answers, and these items received at most two manual answers. The type of manual answer was also variable: a few of them were more general or ambiguous connectives (such as *however* and *but*), some seemed to be related to the syntax of the items (for example, *as of* and *in which*), while others aimed at attributing information between the two arguments to a speaker (for example, *saying* and *stating*). No clear conclusions can be drawn from these insertions.

An additional 1% of the data consisted of 'blank insertions': subjects used the 'manual answer' option to not insert anything. As with the manual answers, there was no clear pattern.

## 6.5 Discussion and conclusion

The current chapter addressed the question of whether a crowdsourced connective insertion task can be used to obtain interpretable discourse annotations, and whether the presence of context influences the resulting annotations.

**Reliability of the task** We found that the method is reliable for acquiring discourse annotations: the majority of inserted connectives converged with the original label, and this convergence was almost perfectly replicable, in the sense that a similar pattern was found in both conditions. The results also showed that subjects often converged on two types of insertions. This indicates that multiple interpretations are possible for a single relation.

Based on these results, we argue that annotation by more than two annotators is necessary, because more observations for every item provides researchers with a probability distribution of all of the senses. This probability distribution will better reflect the true meaning of a relation compared to a single label that is assigned by an annotator using a specific framework.

It should be noted that the design of the method was simplified compared to traditional annotation tasks. First, all items were pre-annotated using one of six classes. In other words, participants were not presented with segments between no relation held (similar to PDTB's NoRel), or with items belonging to a different class than those under investigation (for example, Conditional relations). Second, participants were presented with connectives marking only the six classes under investigation. Including more classes would mean including more connectives. This could could result in less agreement. In the future, we will therefore investigate whether other relational categories (including NoRel) can also be annotated reliably by naïve coders.

The choice of connectives also deserves more consideration. The connectives were chosen to be as unambiguous as possible regarding the six relational classes. The results showed that participants often inserted the connective that corresponded to the original PDTB label. However, for Conjunction relations, causal connectives were also frequently inserted; and for Contrast relations, concessive (negative causal) connectives were often inserted. This indicates that participants correctly interpreted the polarity of the relation, but not the basic operation. This could be due to the connective choice: *in addition* and *by contrast* are less general compared to their causal counterparts, and might therefore be less popular for participants to use. The frequency of connectives might therefore have influenced the readiness of participants to insert the connective.

Regarding the causal connectives, an additional note should be made. Because the PDTB class Cause can contain both objective and subjective causal relations, the connective insertion studies reported in this thesis contained causal connectives that were considered to be ambiguous regarding their source of coherence. However, the status of *as a result* is not entirely clear: it is often considered to be an objective phrase, but some argue that it can also express argumentative relations (see Cohen, 1987). Because of this ambiguity, we recommend future studies to use a different connective that can more clearly express both objective and subjective relations, such as *thus* or *therefore*. Alternatively (and perhaps an even better option), researchers might consider using multiple connectives that can disambiguate between objective and subjective relations. This would, however, mean that participants would have even more connectives to choose from.

**Effect of context on discourse interpretations** The results showed that the presence of context influenced participants' interpretations when the passages contained (at least one of) the following characteristics: (i) the context introduced the topic, (ii) the context sentence following the relation expands on the second argument of the relation; or (iii) the first argument of the relation is a subordinate clause that attaches to the context. The presence of context led to less agreement when the connection between the context and the first argument was not strong due to a paragraph break or a topic change.

However, this only concerned a small portion of the items. Overall, the findings do not support the general consensus that context is necessary for reliable discourse annotation. This suggests that context might not influence the construction of coherence relations during normal reading as much as is assumed in the literature. The lack of a clear effect of context might be attributed to a general ambiguity of language. As Spooren & Degand (2010) note, "establishing a coherence relation in a particular instance requires the use of contextual information, which in itself can be interpreted in multiple ways and hence is a source of disagreement." We believe that the effect of context on readers' interpretations of discourse deserves more consideration and investigation, and therefore presents an interesting area for future research: to what extent do readers take into account the context during discourse interpretation and annotation? What relational "cues" often occur in the context, and how sensitive are readers and coders to such cues? Future studies that aim to answer these questions will hopefully provide more insight into the nature of coherence relations and mental representations.

Despite the lack of a strong effect of context, we do suggest to include context in coherence relation annotation studies if time and resources permit it. When the fragments are presented in their original formatting, context does not have a negative impact on the annotation output, and the presence of context might actually facilitate the inference of a relation, thereby helping annotators who are doubting between multiple senses.

**Methodological remarks on crowdsourcing coherence relation annotations** Most coherence relation annotation efforts in the field are done by a small set of expert annotators. Agreement on annotations can be low (within and between frameworks), even after annotators have received a lot of training. The procedures of traditional relation annotation efforts often lead to implicit biases, which are implemented to achieve a higher amount of annotator agreement (as discussed in Chapter 4). However, from a linguistic or machine learning perspective, annotations containing biases are less useful, because relevant information about additional senses are obscured. Asking a single, expert coder to annotate several readings for a relation also does not solve the issue: annotations would still depend on the knowledge of experts and the annotation process would be more time-consuming. The results reported in this chapter indicate that crowdsourcing can provide a solution.

The method that is proposed is – at a general level – similar to PDTB's annotation process: annotators are required to insert a connective that marks the relation between the segments. However, there are several crucial differences between the methodologies, such as the use of naïve, untrained individuals in our study, the lack of an annotation stage comparable to the PDTB approach (that is, when PDTB annotators assign a specific label to the relation after choosing a connective), and the larger number of observations in our study. Moreover, our participants could only choose from a small set of connectives, and these connectives are less ambiguous compared to many of the markers that PDTB annotators could insert. Depending on the condition, the participants also had only a few or no context sentences available, in contrast to PDTB annotators, who are provided with the entire text.

The most crucial difference between the method described in the current chapter and traditional annotation methods is the resulting data. Inserting connectives does lead to more coarse-grained annotations when compared to the relation label annotations of trained annotators. However, annotations obtained in the current study better reflect the interpretations of average readers because naïve annotators do not rely on biases and rules such as those inherent in annotation frameworks. Moreover, crowdsourcing annotations is easier, more affordable and faster than traditional annotation methods. Collecting a large number of observations for a single item furthermore provides us with a distribution of senses, which can be useful for future efforts investigating the nature of multi-interpretable relations.

However, there are also limits in interpretability due to the experimental design. For example, based on our results, it is not possible to decide whether relations that received multiple types of insertions were genuinely ambiguous to every participant (i.e. the participants inferred both readings but decided for expressing only one of the readings) or whether different participants had different interpretations of a single relation (i.e. every participant could only infer one of the readings). The participants were provided with the option of inserting two connectives for every item, but they barely made use of this option. One possibility is that participants did not provide double insertions because they only had a single reading of the item. However, motivation might also have played a role in the lack of double insertions. The task required participants to insert at least one connective (not necessarily more). Because inserting a second connective requires more time, participants might have chosen to insert only one, even if they inferred multiple readings for a relation. For future experiments focusing on possible double senses of relations, this issue can be solved by making it obligatory to provide two observations per items; in other words, participants would need to indicate explicitly if they cannot infer a second reading (the option of “No other connective fits"). Note that a similar approach is taken in the construction of PDTB 3.0 (see Webber et al., 2016).

Finally, we note that some items received many very different annotations. This might be due to participants’ lack of domain knowledge in economics. Such a lack of domain knowledge likely not only affects participants recruited via crowdsourcing platforms, but may also affect the annotations of trained annotators who may not be very familiar with the textual domain. It is possible that, as a community, we underestimate the effect of familiarity with a domain on the quality and reliability of coherence relation annotations.

In sum, based on the near-identical patterns of sense distributions in two conditions, we conclude that there is merit in the crowdsourced connective insertion method, and believe that with some adaptations, it can be used to create a corpus. Kawahara et al. (2014) and Rohde et al. (2016) have put forward related approaches using crowdsourced relation annotation tasks. These studies also advocate crowdsourcing, because annotations are obtained quickly and affordably, the resulting data are reliable, and it can provide valuable insights that traditional annotation tasks cannot (at least not as easily). What sets the current method apart from previous work is its comparability to PDTB’s annotations, given that the connectives were chosen to match PDTB’s classes. In its current state, the method leads to more coarse-grained annotations compared to PDTB, but it is conceivable that it can be extended to reflect more fine-grained distinctions in interpretations. Future research will hopefully focus on which connectives can be added to represent more distinctions (e.g., Temporal relations), as well as how the lower agreement for Conjunction and Contrast relations can be improved. If these issues are solved, the connective insertion task has the potential to be used for creating a discourse-annotated corpus that embraces multiple interpretations. The method can also prove to be useful for answering other types of research questions. In the next two chapters, readers’ interpretations of ambiguous coherence relations are investigated by studying the distribution of insertions obtained via this crowdsourcing method.

# Chapter 7

# Identifying elaborative and argumentative discourse relations[1]

Chapter 4 revealed that PDTB and RST-DT annotators often disagree on the annotation of examples and specification relations; specifically, there is disagreement about whether they are ideational (additive) or argumentative (pragmatic causal). Chapter 5 argued that in order to evaluate such discrepancies in annotations, we should consider evidence from various sources, including from interpretation studies. Examples and specifications have been the topic of many theoretical studies, but no research has investigated how readers interpret them. The current chapter aims to evaluate the discrepancies in annotations by presenting a crowdsourced connective insertion study, which elicited the interpretation of these relations from naïve readers.

The results show that these relations can indeed have two functions: they can be used to illustrate / specify a situation and to serve as an argument for a claim. These findings suggest that examples and specifications can have multiple, simultaneous readings. This contributes to the discussion regarding the multi-level thesis: previous theoretical work had argued that relations can function on different levels, but that one reading would be most prevalent for every relation. The current study is the first to investigate how readers interpret these different levels, and the results indicate that the readings corresponding to the different levels are in fact both prevalent for many items. We discuss the implications of these results for discourse annotation.

[1] This chapter is adapted from, and in some parts identical with, Scholman, M.C.J. and Demberg, V. (2017). Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue & Discourse,* 8(2), 56–83.

## 7.1 Introduction

Discourse-annotated corpora allow researchers to conduct quantitative investigations of coherence relations. The largest and most frequently used corpora are the Penn Discourse Treebank (PDTB; Prasad et al., 2008) and the Rhetorical Structure Theory Discourse Treebank (RST-DT; Carlson, Marcu & Okurowski, 2003). In the study reported in Chapter 4, annotations from overlapping texts in these corpora were mapped onto each other in order to determine whether annotations from these corpora are compatible (that is, whether the frameworks agree on the sense of a relation). The results showed that inter-framework agreement on explicit relations was reasonable (ca. 60% agreement), but agreement on implicit relations was much lower (approximately 35% agreement). An obvious explanation for this difference in agreement is the presence of connectives: achieving high inter-annotator agreement on implicit relations is more difficult than on explicit relations because annotators cannot rely on the information provided by connectives.

Even when taking into account the difficulty of annotating implicit coherence relations, the amount of agreement between frameworks on the same texts seems low. In this chapter, we focus on two types of relations for which inter-framework agreement is particularly poor, and investigate the factors affecting the interpretation of these relations in more detail. The goal is to establish whether the different annotations are justified by looking at naïve readers' interpretations. The relations under investigation are PDTB's INSTANTIATION and SPECIFICATION relations (32% and 14% agreement, respectively). These relations do not have many corresponding prototypical connectives (Taboada & Das, 2013; Vergez-Couret & Adam, 2012) and are very frequently implicit: together, they make up 24% of all implicit relations in the PDTB. Both types of relations are subtypes of the PDTB class EXPANSION. INSTANTIATION is a second-level label in the PDTB hierarchy, whereas SPECIFICATION is a third-level label belonging to the PDTB type RESTATEMENT (see Appendix A for the PDTB and RST-DT taxonomies). For convenience, EXPANSION.RESTATEMENT.SPECIFICATION is referred to as SPECIFICATION in this thesis. Both of these relation types are characterized by one segment further specifying a set or situation described in the other segment (Halliday, 1994).

INSTANTIATIONS and SPECIFICATIONS are considered as primarily elaborative (also referred to as additive) relations. A large proportion of the mismatch with the corresponding RST-DT annotations stems from these same relations being annotated as argumentative (causal); namely EXPLANATION-ARGUMENTATIVE and EVIDENCE. In such argumentative relations, one segment gives support to the other segment (Jasinskaja & Karagjosova, 2011).

As Blakemore (1997) argues, INSTANTIATIONS and SPECIFICATIONS can indeed have two functions: they can be used to illustrate / specify a situation and to serve as an argument to a claim. PDTB allows the annotation of more than one relation sense, but the argumentative function of INSTANTIATION and SPECIFICATION relations is usually not annotated. PDTB rather focuses on the illustrative / specification aspect of the relation. However, if readers do infer the argumentative reading of the INSTANTIATIONS and SPECIFICATIONS, it is important that this reading is also reflected in a corpus. Otherwise, the annotation of those INSTANTIATIONS and SPECIFICATIONS cannot be considered as fully descriptively adequate and cognitively plausible. Similarly, when RST-DT annotates the argumentative aspect of the relation (by labeling relations as EXPLANATION-ARGUMENTATIVE and EVIDENCE), the illustrative / specification aspect of the relation is not reflected in the annotation (which is reflected by the RST-DT labels EXAMPLE and GENERAL-SPECIFIC).

Considering that INSTANTIATION and SPECIFICATION are so frequent, the current study sets out to investigate how readers actually interpret these relations. Do PDTB's INSTANTIATION and SPECIFICATION relations vary in the degree to which they can be interpreted as argumentative? Can we identify specific characteristics of the relations that have an additional argumentative function? Are there also other alternative interpretations that are inferred?

We aim to answer these questions by asking crowdsourced participants to insert a connective from a predefined list between the segments of coherence relations. The results can provide a better understanding of the functions of INSTANTIATION and SPECIFICATION relations, and, for annotation purposes, to more reliably identify and classify INSTANTIATION and SPECIFICATION relations, which in turn can improve the agreement on these frequently-occurring classes.

In sum, the current chapter deals with the following issues: (i) the mismatch between PDTB INSTANTIATION and SPECIFICATION relations with RST-DT's annotations, and (ii) the variability in presumably valid interpretations of INSTANTIATION and SPECIFICATION relations (elaborative vs. argumentative). The layout of the chapter is as follows: in the next section, we discuss INSTANTIATION and SPECIFICATION relations and their signaling. We then present a crowdsourcing experiment, showing that INSTANTIATION and SPECIFICATION relations are indeed often interpreted by our participants as argumentative. The chapter concludes with a discussion of the results and their implications for discourse annotation.

## 7.2 Background

INSTANTIATION and SPECIFICATION belong to the PDTB class EXPANSION. The relation types are quite similar. A relation is labeled as INSTANTIATION when "the connective indicates that Arg1 evokes a set and Arg2 describes it in further detail" (Prasad et al., 2007, p. 34).[2] The set that is described in Arg1 may be a set of events, reasons, behaviors, etc. Typical markers of INSTANTIATIONS include *for example* and

[2] PDTB uses a lexical grounding approach, for which connectives are annotated for both explicit and implicit relations. In the case of implicit relations, annotators insert a connective and annotate the corresponding sense of the relation.

*as an illustration.* Example (23) illustrates this type of relation. In SPECIFICATION relations, Arg2 also describes Arg1 in more detail, but Arg2 is also a logical implication of Arg1. Typical markers of SPECIFICATION relations include *specifically* and *in fact.* Example (24) presents an example of a SPECIFICATION relation. The relation types are therefore similar in that Arg2 expands on Arg1, but they differ in the presence or absence of a set and logical implication.

(23) In an age of specialization, *the federal judiciary is one of the last bastions of the generalist.* **A judge must jump from murder to antitrust cases, from arson to securities fraud, without missing a beat.**[3] wsj_601

(24) *In the cornucopia of go-go apples, the Fuji's track record stands out.* **During the past 15 years, it has gone from almost zilch to some 50% of Japan's market.** wsj_1128

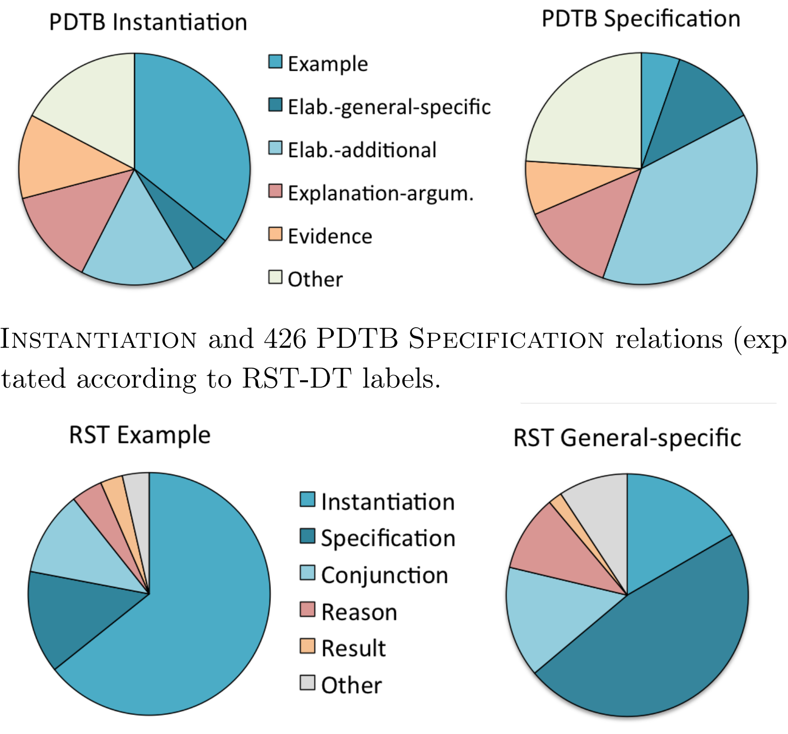
RST-DT distinguishes relation labels that map onto PDTB's INSTANTIATION and SPECIFICATION: EXAMPLE for INSTANTIATION and ELABORATION-GENERAL-SPECIFIC[4] for SPECIFICATION. However, the mapping study reported in Chapter 4 showed that PDTB's INSTANTIATION and SPECIFICATION relations often fall into other classes in RST-DT. As illustrated in Figure 7.1a, other common labels for INSTANTIATION relations are GENERAL-SPECIFIC (which maps onto PDTB's SPECIFICATION label), ELABORATION-ADDITIONAL (the most basic relation in RST-DT), EXPLANATION-ARGUMENTATIVE, and EVIDENCE (both typically causal labels in RST-DT). These same RST-DT labels were used for relations that received the SPECIFICATION label in PDTB. Hence, RST-DT often assigns causal labels to relations that PDTB labels as elaborative INSTANTIATION or SPECIFICATION (25% of PDTB INSTANTIATION and 21% of PDTB SPECIFICATION relations receive causal labels in RST-DT).

RST-DT's classes EXAMPLE and GENERAL-SPECIFIC are not characterized by the same pattern: Figure 7.1b shows that PDTB annotators did not assign a causal label as often to RST-DT's EXAMPLE and GENERAL-SPECIFIC relations (7% and 12%, respectively). Hence, for relations that RST-DT annotators consider to be EXAMPLES and GENERAL-SPECIFICATIONS, PDTB annotators tend to annotate the same reading.

This mismatch regarding the annotation of INSTANTIATIONS and SPECIFICATIONS could be attributed to the frameworks' procedures. PDTB has a connective-based approach: annotators are instructed to annotate the connective if one is present, or insert one and then annotate the relation, when the relation is implicit. Annotators are not instructed to systematically try to insert certain connectives before trying others.

[3] In line with the PDTB, the first argument of a relation (Arg1) is indicated with italic font and the second argument (Arg2) with bold font.

[4] RST-DT's ELABORATION-GENERAL-SPECIFIC will be abbreviated to GENERAL-SPECIFIC in the remainder of this chapter.



**(a)** 306 PDTB Instantiation and 426 PDTB Specification relations (explicit and implicit) annotated according to RST-DT labels.

**(b)** 168 RST-DT Example and 108 RST-DT General-specific relations (explicit and implicit) annotated according to PDTB labels.

**Figure 7.1:** Mappings for PDTB Instantiation and Specification instances and RST-DT Example and General-specific instances from the aligned corpus (see Chapter 4). Relational labels corresponding to blue colors are primarily elaborative, orange labels are primarily causal.

They can choose from a list containing many connectives, and multiple connectives can often fit a single relation. For example, PDTB annotators inserted the discourse marker *specifically* for Example (24) above, but *because* could have also been inserted in this relation. The choice of inserted connective hence depends on the annotator, and this can vary from annotator to annotator. It might be that different annotators have different biases for inferring one sense over another, but this is not retraceable in the corpus itself. It is therefore difficult to determine what the framework's bias is.

By contrast, RST-DT annotators do not explicitly make use of connectives at all; instead, the framework instructs annotators to focus on the writer's intentions. From this perspective, it is likely that even though *specifically* or *for example* can express the relation, the intention of the writer has a more argumentative nature: to convince the reader of a claim by providing evidence. Indeed, PDTB's Instantiation and Specification classes contain instances that can be analyzed as consisting of a claim and an argument. In the case of Instantiation relations, the second segment can be seen as supporting a claim in the first segment. Consider the following example, annotated as PDTB INSTANTIATION – RST-DT EVIDENCE:

(25) *"If you are born to give parties, you give parties.* **Even in Russia we managed to give parties."** wsj_1367

The first segment of this example is a claim, and the second segment gives an illustration of the claim. This reading is referred to as the ideational (or semantic) reading, which involves the relation between the information conveyed in consecutive elements of a coherent discourse (cf. Moore & Pollack, 1992). However, the second segment can also be interpreted as a premise that underpins the validity of the claim. From this perspective, the relation is in fact pragmatic causal, or argumentative.[5] Argumentative relations are relations in which the writer attempts to affect the addressee's beliefs, attitudes, or desires by means of language (cf. Hovy & Maier, 1995).

SPECIFICATION relations can be similarly ambiguous between an ideational and an argumentative reading: the second segment can serve as providing more information about a concept or situation in the first segment, or it can provide support for a claim in the first segment. Carston (1993, p. 164) also addressed this double function by noting that "exemplification is a common way of providing evidence to support a claim, or, equivalently, of giving a reason for believing something." Building on this, Blakemore (1997) argued that INSTANTIATION and SPECIFICATION relations can have different functions in a text, and that classifying them as either ideational or argumentative does not do justice to these relations true interpretations. The double function of INSTANTIATION and SPECIFICATION relations has also been noted in other descriptive work on ELABORATION relations (see, for example, Cuenca, 2003; Hyland, 2007; Jasinskaja & Karagjosova, 2011), but it is not reflected in coherence relation annotation frameworks. The current study builds on the assumption that the classification of INSTANTIATIONS and SPECIFICATIONS is important, considering that corpora should be descriptively adequate and cognitively plausible.

We can conclude (based on Figures 7.1a and 7.1b) that PDTB's INSTANTIATION and SPECIFICATION classes and RST-DT's GENERAL-SPECIFIC class contain both argumentative and non-argumentative relations. This finding is in fact not surprising: Peldszus & Stede (2013) and Stab & Gurevych (2014) noted that these frameworks do not have distinct labels for specific argumentative relations since they are focused on identifying general discourse structures, and the PDTB focuses on semantic relations explicitly. Biran & Rambow (2011) also note that argumentation is not characterized by a single coherence relation type; rather, it can be realized by a large number of coherence relation types. This does not imply that all relation types have an argumentative and a non-argumentative reading, but it does apply to INSTANTIATION and SPECIFICATION relations. If readers can actually infer an argumentative reading

[5] The terms ideational and argumentative are also known as informational and intentional (Moore & Pollack, 1992), subject matter and presentational (Mann & Thompson, 1988), and semantic and interpersonal (Hovy & Maier, 1995).

for such elaborative relation types, annotation frameworks might need to be able to represent the argumentative reading of those non-causal relations.

The argumentative reading of a non-causal relation can be incorporated in its annotation by annotating multiple senses for the same relation. Consider Example (25): such relations actually have multiple readings (both elaborative and argumentative), and multiple labels would therefore adequately express the readings of this relation. However, most annotated relations that are currently available do not carry double annotations,[6] because these corpora are built on the assumption that a *single* coherence relation holds between two segments. This assumption has been challenged in recent years. For example, investigations of the discourse adverbial *instead* have shown that some relations can actually express multiple readings (Webber, 2013). *Instead* prototypically marks Exception relations, but when *instead* occurs at the beginning of a segment, another reading can often be inferred as well, as illustrated in (26)-(28), taken from Rohde et al. (2016).

(26) *I planned to make lasagna.* Instead **I made hamburgers.**
→ But instead **I made hamburgers.**

(27) *I don't know how to make lasagna.* Instead **I made hamburgers.**
→ So instead **I made hamburgers.**

(28) *Surprisingly, they ignored the lasagna.* Instead **they just ate the salad.**
→ And instead **they just ate the salad.**

As Examples (26)-(28) show, relations marked by *instead* can express different readings, including the Exception relation that *instead* signals. Building on this observation, Rohde et al. (2016) showed that relations containing discourse adverbials other than *instead* can also express multiple readings. These results indicate that multiple, simultaneous coherence relations can hold between two discourse segments that are connected by an adverbial. Based on these findings, we can expect Instantiation and Specification relations to also have multiple readings.

In sum, the fact that annotators often disagree on the elaborative or argumentative nature of PDTB's Instantiation and Specification relations can be due to an ambiguity or possible double-function of these relations that is not captured in the frameworks. The current study therefore investigates how comprehenders interpret these relations. The results will provide insight into the validity of the annotations and the distinctions between these relations. The obtained annotations will also be used to identify certain cues that might have influenced the readers' interpretations (in other words, cues that signal the elaborative or argumentative type of Instantiation and Specification relations).

[6] PDTB does allow for double annotations, but this is rarely applied in practice: less than 5% of instances in the PDTB carry two labels. The PDTB group plans to release a new version, PDTB 3.0, in which double labels occur more frequently (Webber et al., 2016).

## 7.3 Method

We conducted a crowdsourcing experiment for which naïve (non-trained, non-expert) annotators were asked to insert connectives from a predefined list into coherence relations taken from the mapped Wall Street Journal texts. The data comes from the context condition in the study reported in Chapter 6 (but note that the no-context condition showed a perfect replication of the results). This section presents information on the method that is relevant to the research goals of this study; for more details regarding the general design, see Chapter 6.

### 7.3.1 Participants

111 native English speakers (age range 22-68; mean age 36 years; 47 female) completed one or more batches of this experiment. They were recruited via Prolific and reimbursed for their participation (2 GBP per batch). Participants came from the United States, United Kingdom, Ireland, and Australia. Their education level ranged from an undergraduate degree to a doctorate degree.

### 7.3.2 Materials

The experimental passages were implicit INSTANTIATION and SPECIFICATION relations taken from the aligned corpus reported in Chapter 4. These relations were chosen to enable a comparison between the PDTB label and the RST-DT label. The INSTANTIATION and SPECIFICATION relations that were included in this experiment fell into one of four RST-DT classes: EXAMPLE, GENERAL-SPECIFIC, EVIDENCE, and EXPLANATION-ARGUMENTATIVE. These categories were chosen because relations that received PDTB's INSTANTIATION or SPECIFICATION label fall in these four RST-DT classes most often. In total, 54 INSTANTIATION and 60 SPECIFICATION items were included in this experiment. For each of the four relevant RST-DT classes, 15 INSTANTIATION and 15 SPECIFICATION items were taken, with the exception of the INSTANTIATION – GENERAL-SPECIFIC combination: there were only 9 items the PDTB label INSTANTIATION and the RST-DT label GENERAL-SPECIFIC. A list of item identifiers is provided in Appendix F.

When the aligned corpus contained more than 15 items annotated with the target PDTB and RST-DT label, preference was given to items that (i) differed less than 60 characters between the RST-DT and PDTB segmentation, (ii) did not contain attribution in one of the arguments, and (iii) dealt with a non-economic topic. The first criterion relates to the size of the segments: PDTB and RST-DT have different segmentation rules, resulting in different segment sizes. We chose relations that differed as little as possible in segmentation. We adhered to the PDTB segmentation (rather than RST-DT) because PDTB annotates the minimal amount of information necessary to infer the intended relation. The second criterion relates to attribution in the segments; that is, the explicit reference to the source (e.g., *John said that*). PDTB does not annotate attribution as part of an argument, but rather as a feature of the relation. Therefore, if the source of the attribution was reported in between the two arguments, it was moved to the context sentence following the second argument, to ensure that participants did not treat it as part of the argument. The third criterion was composed for the motivation of participants: non-economical topics were considered more interesting to read than economical topics.

The set of fillers consisted of 24 Cause, 24 Conjunction, 36 Concession and 36 Contrast relations. To ensure that the fillers in this experiment were clear cases of a specific type of relation, they were selected based on the criterion that the PDTB and RST-DT label were in agreement.

The total of 234 items was divided into 12 batches, with 4 or 5 Instantiation, 5 Specification, 2 Cause, 2 Conjunction, 3 Concession, 3 Contrast items per batch. Order of presentation of the items per batch was randomized to prevent order effects. Participants were allowed to complete more than one batch, but saw every item only once. Average completion time per batch was 16 minutes. Due to presentation errors in one Conjunction, two Cause, and two Concession items, the final dataset for analysis including fillers consists of 229 items.

**Connecting phrases** The list of connecting phrases was identical to that used in Chapter 6 and consisted of: *as an illustration*, *more specifically*, *in addition*, *because*, *as a result*, *even though*, *nevertheless*, and *by contrast.*

### 7.3.3 Procedure

The experiment was distributed via Prolific and hosted on LingoTurk (Pusse, Sayeed & Demberg, 2016). The method is identical to the method reported in Chapter 6. All items were presented with two context sentences preceding and one context sentence following the segments.

## 7.4 Results

Prior to analysis, the data of 4 participants were removed because these participants had very short completion times (<10 minutes for 20 passages of 5 sentences each) and showed high disagreement on causal and concessive items with other participants. The following analyses do not take the responses of these participants into consideration, leaving us with a total of 2962 observations. In total, each list was completed by 12 to 14 participants.

As with any discourse annotation task, some variation in the distribution of insertions can be expected. We are therefore interested in larger patterns in the distribution of inserted connectives. Crucially, we do not assume that there is one single

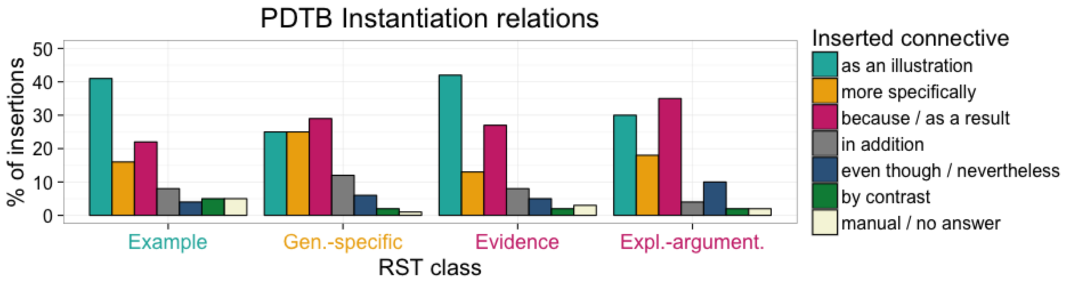


**Figure 7.2:** Distribution (%) of inserted connectives in INSTANTIATION relations per RST-DT class.

correct label for each of our experimental items (cf. Rohde et al., 2016). In cases where we observe a distribution of multiple connectives for a single item, we take this to indicate that this particular relation might have multiple readings.

For the following analyses, we aggregated frequencies of the connectives that fell into the same class. In other words, *because* and *as a result* were aggregated as causal connectives, and *even though* and *nevertheless* were aggregated as concessive connectives.

First, we analyze insertions into PDTB INSTANTIATIONS (Section 7.4.1). We will then look at the insertions per RST-DT class (Section 7.4.1) to investigate whether the subjects interpreted the items in line with PDTB's or RST-DT's classification. Next, we will look at a few examples of clear elaborative and clear argumentative relations in more detail to be able to identify cues for these relation types (Section 7.4.1). The same is discussed for SPECIFICATION items (Sections 7.4.2 and 7.4.2).

## 7.4.1 Analysis of INSTANTIATION relations

In this section, we look at the insertions into INSTANTIATION relations, first by RST-DT label (Section 7.4.1) and then by item (Section 7.4.1).

### Analysis of INSTANTIATION relations by RST-DT label

Figure 7.2 shows the distribution of inserted connectives in INSTANTIATION relations per RST-DT class. This visualization allows us to investigate whether the insertions that we obtained in the experiment agree with the reading that is annotated by RST-DT annotators (elaborative or argumentative).

The first RST-DT class under investigation, EXAMPLE, is considered to be similar to PDTB's INSTANTIATION. As these two labels are direct correspondences to one another, we predicted that the participants would also agree with this label and hence most frequently choose the connective *as an illustration*. As Figure 7.2 shows, the instantiation marker *as an illustration* was indeed the most frequent connective chosen (41% of all insertions), but the distribution is quite broad: there were also many causal connectives (22% of inserted connectives) and many Specification insertions (16% of insertions).

The second class, General-specific, is considered to be inconsistent with PDTB's class Instantiation, since PDTB's Instantiation corresponds to RST-DT's Example, and PDTB's Specification corresponds to General-specific. We predicted that these relations might be genuinely ambiguous between an Instantiation or a Specification reading, and that we would see both Instantiation and Specification markers inserted. The results confirm this prediction: items in the General-specific class received an equal number of Instantiation and Specification insertions (both 25%). However, the most frequently inserted type of connectives was causal, taking up 29% of all inserted connectives in this class, and we also see more Conjunction relations than for other subgroups of Instantiation relations. This group of relations therefore seems to be quite ambiguous.

The third RST-DT class under investigation, Evidence, is generally considered to be a causal class. We therefore predicted that these relations may have two functions: an example that also serves as evidence for a claim. We expected that both Instantiation and causal markers would both be inserted often. Figure 7.2 confirms this prediction: the Instantiation marker was inserted most often (42% of all insertions), with causal connectives as the second most frequently inserted type (27%).

Finally, for the items annotated as PDTB Instantiation and RST-DT Explanation-argumentative, it was also expected that both Instantiation and Cause markers would be inserted. This was indeed the case: 30% of inserted markers were *as an illustration*, and 35% were *because / as a result*. Furthermore, participants often inserted connectives expressing Concession relations for this subgroup, which may reflect the causal aspect of these relations (Concessions are negative causal relations). Again, the Specification marker was also inserted relatively often, accounting for 18% of all insertions.

In sum, we find that all of the subclasses had a substantial number of Instantiation, Specification and Cause interpretations. The results from our study on PDTB Instantiation relations could be interpreted as evidence that Instantiation items have both an elaborative *and* an argumentative function. However, it is also possible that rather than items being complex or ambiguous, subjects interpret a proportion of the Instantiation items as expressing an elaborative relation, and another proportion presenting an argumentative relation. The next section provides more insight into this issue.

### By-item insertions in Instantiation items

For the analyses in the previous section, all Instantiation items were grouped together and the percentages represented an average number of insertions (over all items) per RST-DT class. This grouping obscures any possible differences between

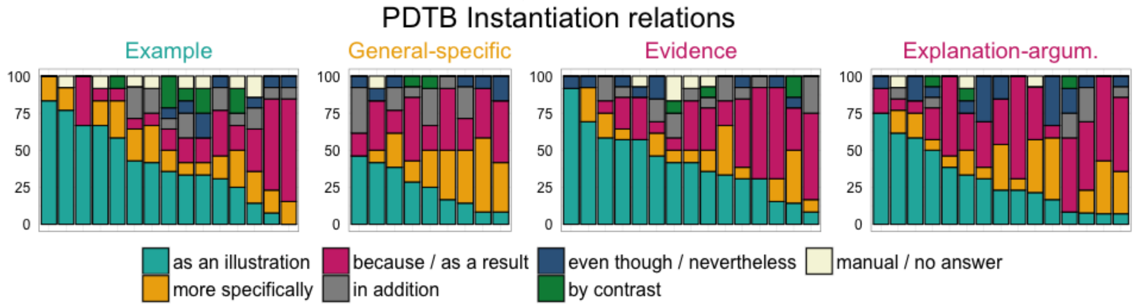


**Figure 7.3:** Distribution (%) of inserted connectives in Instantiation relations per RST-DT class and item. Plots are arranged according to the number of Instantiation insertions.

| **Relation type** | **Agr. with PDTB** | **Agr. with RST-DT** |
|---|---|---|
| Instantiation – Example | 73 | 73 |
| Instantiation – General-specific | 33 | 22 |
| Instantiation – Evidence | 67 | 27 |
| Instantiation – Explanation-argum. | 33 | 60 |

**Table 7.1:** Percentage agreement between the dominant response and the PDTB label Instantiation, per RST-DT class.

items of the same RST-DT class. In this section, we look at the distribution of insertions per item. The distribution per item can be used to observe genuine ambiguity in the interpretation of some items, but also to derive the dominant interpretation for each item.

Figure 7.3 provides a detailed picture of the percentage of inserted connectives per RST-DT class and item. Every stacked bar on the x-axis represents an item, and the colours on the bars represent the inserted connectives. One way to analyze this data is to assign to each relation the label corresponding to the connective that was inserted most frequently by our participants, referred to as the *dominant response* (in Figure 7.3, this corresponds to the largest bar per item). After assigning relations the label corresponding to the dominant response, we can calculate how many items received a dominant response identical to the PDTB or RST-DT label. In other words, we can calculate agreement between the dominant response per item and the PDTB label and RST-DT label. This result is reported in Table 7.1.

Table 7.1 shows that the dominant response converges with the PDTB label relatively often for items in the RST-DT classes Example and Evidence. For items in the class Instantiation – General-specific, the dominant response does not converge with the PDTB label more, or less, often than with the RST-DT label. This supports the hypothesis that these relations are ambiguous. The other common dominant response for these Instantiation – General-specific items is causal. For items in the class Instantiation – Explanation-argumentative, the dominant response is most often causal, thereby converging with the RST-DT label.

The visualization of the by-item analysis in Figure 7.3 reveals some trends that we will discuss in more detail. Items that elicited many Instantiation insertions mainly belong to the RST-DT classes Example and Evidence, with a few cases also occurring in the class Explanation-argumentative. These items are considered clear examples of the class Instantiation, and we expect there to be a cue present that indicates to readers that the item is an Instantiation relation. A closer look at these items revealed a common characteristic: often, a larger set is mentioned in the first argument, and one member of the set is explicitly referred to in the second argument. This larger set is referred to by a quantifier such as 'many', by plural noun phrases such as 'glossy brochures' and 'larger department stores', or by a combination of a quantifier and a plural noun phrase. This is illustrated in example (29), which is taken from the Instantiation – Example class and is presented in the same way as it was presented to participants. In Arg1 of Example (29), the set 'glossy brochures' is mentioned. Arg2 then refers to one member of the set ('one handout') and gives a more specific example of the phenomenon described in Arg1.

(29) But that's for the best horses, with most selling for much less – as little as $100 for some pedestrian thoroughbreds. Even while they move outside their traditional tony circle, racehorse owners still try to capitalize on the elan of the sport.
*Glossy brochures circulated at racetracks gush about the limelight of the winner's circle and high-society schmoozing* // **one handout promises: Pedigrees, parties, post times, parimutuels and pageantry.**
"It's just a matter of marketing and promoting ourselves," says Headley Bell, a fifth-generation horse breeder from Lexington. wsj_1174

The items that elicited mainly causal insertions occurred predominantly in the RST-DT classes Evidence and Explanation-argumentative (with two items in the class Example). A common trait of these causal Instantiations is that the first segment consists of a subjective utterance that can be interpreted as a claim and the second segment contains an argument for this claim, as in Example (30), taken from the group Instantiation – Evidence: The speaker makes a claim in the first segment, and provides evidence for this claim in the second segment, as well as in the context following the second segment. The majority of the subjects interpreted this relation as causal (62%).

(30) That done, Ms. Volokh spoke with rampant eloquence about the many attributes she feels she was born with: an understanding of food, business, Russian culture, human nature, and parties. "Parties are rather a state of mind," she said, pausing only to taste and pass judgment on the Georgian shashlik ("a little well done, but very good").

*"If you are born to give parties, you give parties //* **even in Russia we managed to give parties.**
In Los Angeles, in our lean years, we gave parties." wsj_1367

Another characteristic of relations that elicited causal insertions is that the second segment can be interpreted as a result of the situation described in Arg1, as in Example (31), taken from the group Instantiation – Example. In this example, the Instantiation reading can be inferred when readers interpret the second segment as an example of how international competition is heating up. However, when readers interpret these segments as occurring chronologically, they will get the reading that the situation in the second segment happens as a result of the situation in the first segment. Indeed, 54% of insertions in this item were *as a result* (and 15% of insertions were because).

(31) The goal of most U.S. firms – joint ventures – remains elusive. Because the Soviet ruble isn't convertible into dollars, marks and other Western currencies, companies that hope to set up production facilities here must either export some of the goods to earn hard currency or find Soviet goods they can take in a counter-trade transaction.
*International competition for the few Soviet goods that can be sold on world markets is heating up, however //* **West German companies already have snapped up much of the production of these items.**
Seeking to overcome the currency problems, Mr. Giffen's American Trade Consortium, which comprises Chevron Corp., RJR, Johnson & Johnson, Eastman Kodak Co., and Archer-Daniels-Midland Co., has concocted an elaborate scheme to share out dollar earnings, largely from the revenues of a planned Chevron oil project. wsj_1368

Finally, certain items received many different types of insertions without showing a clear dominant response. Manual inspection revealed that these items often revolve around topics of economics that typically require background knowledge about the stock markets. The lack of agreement in annotation of these relations may hence be due to participants not having enough background information to judge the relations in the text. Given that even professionally trained discourse relation annotators are often not experts on the topic of the text that is being annotated, it is possible that this domain problem also affects the original PDTB and RST-DT-DT annotations (see also Martins, Kigiel & Jhean-Larose, 2006; McNamara, Kintsch, Songer & Kintsch, 1996).

### 7.4.2 Analysis of Specification relations

In this section, we look at the insertions into Specification relations, first by RST-DT label and then by item.

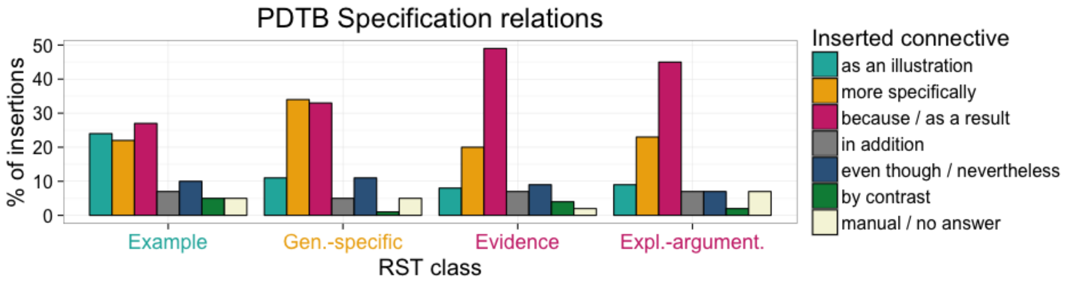


**Figure 7.4:** Distribution (%) of inserted connectives in Specification relations per RST-DT class.

### Analysis of Specification relations by RST-DT label

For items from the PDTB class Specification, the most frequently inserted connective type was not the marker that would be expected based on the PDTB label, but a causal marker (39%). The connective *more specifically* was the second most frequent type (25%) and *as an illustration* was the third most frequent type (13%). We again split up the dataset by RST-DT labels for a more detailed analysis, see Figure 7.4.

The same predictions that held for Instantiation items per RST-DT class also hold for Specification items: for items annotated as PDTB Specification and RST-DT Example, we predicted that this disagreement between PDTB and RST-DT annotations would be reflected in a similar split of inserted connectives by our participants. As Figure 7.4 shows, this is indeed the case: subjects inserted the Instantiation marker in 24% of the cases, and the Specification marker in 22% of the cases. Somewhat surprisingly, we find a large proportion of causal insertions (27%) for these instances. Similar to the findings in Section 7.4.1, we find that relations for which PDTB and RST-DT annotators did not agree on the Instantiation or Specification sense are ambiguous.

Items annotated as PDTB Specification and RST-DT General-specific received a nearly equal number of Specification and causal insertions (34% and 33%, respectively). This brings up the question of whether these instances are in fact both elaborative and argumentative. This will be discussed in the by-item analysis in Section 7.4.2 below.

For items that received the PDTB Specification and RST-DT Evidence label, we predicted that participants would insert a high number of Specification and Cause markers. As Figure 7 shows, nearly half of all insertions were causal (49%), while only 20% of insertions were *more specifically*. Hence, participants seem to pick up on the same reading as RST-DT annotators for these items. A similar pattern occurs for Specification items that received the RST-DT label Explanation-argumentative: 45% of the insertions were causal, and 23% were *more specifically*.

These results indicate that naïve subjects tend to interpret Specification items as expressing a causal relation. The by-item analysis in the next section will show that items often received both types of insertions, and not only one or the other type.

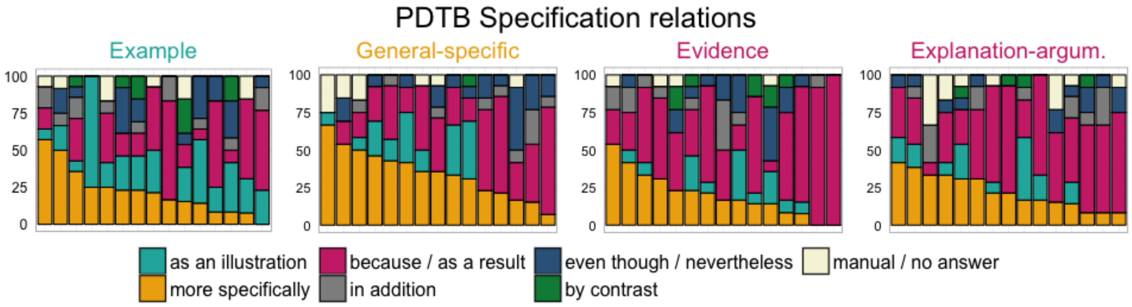


**Figure 7.5:** Distribution (%) of inserted connectives in SPECIFICATION items per RST-DT class and item. Plots are arranged according to the number of SPECIFICATION insertions.

| Relation type | Agr. with PDTB | Agr. with RST-DT |
|---|---|---|
| SPECIFICATION – EXAMPLE | 20 | 27 |
| SPECIFICATION – GENERAL-SPECIFIC | 40 | 40 |
| SPECIFICATION – EVIDENCE | 13 | 73 |
| SPECIFICATION – EXPLANATION-ARGUM. | 27 | 67 |

**Table 7.2:** Percentage agreement between the dominant response and the PDTB label SPECIFICATION, per RST-DT class.

### By-item insertions in SPECIFICATION items

Figure 7.5 displays the distribution of inserted connectives in PDTB SPECIFICATION relations per RST-DT class.

From these distributions of answers per items, we again calculate dominant responses and their agreement with the original PDTB and RST-DT labels. The results of this analysis are shown in Table 7.2. The dominant response only rarely converges with the PDTB label. The most agreement is observed for items on which the PDTB and RST-DT annotations agree (RST-DT GENERAL-SPECIFIC).

A common characteristic of items receiving a high number of SPECIFICATION insertions is that the first segment contains a reference to a general or vague concept, such as ‘one thing’ in Example (32).

(32) The LDP won by a landslide in the last election, in July 1986. But less than two years later, the LDP started to crumble, and dissent rose to unprecedented heights.
*The symptoms all point to one thing* // **Japan does not have a modern government.**
Its government still wants to sit in the driver’s seat, set the speed, step on the gas, apply the brakes and steer, with 120 million people in the back seat.

For items in the class SPECIFICATION – EXAMPLE, we find that the dominant response does not converge with either of the labels very often. For items in the class Specification – Evidence and Specification – Explanation-argumentative, the dominant response is often causal. It thereby converges with the RST-DT label (also shown in Table 7.2). A manual analysis of these items showed a similar finding to that discussed in Section 7.4.1: often, the first segment of the relation contains a subjective claim. It is likely that readers interpreted the more specific information in the second segment as evidence for the claim.

### 7.4.3 Double insertions and manual answers

Participants were given the option of inserting two connecting phrases if they thought that both phrases reflected the meaning of the relation. For Instantiation and Specification items, 4.1% of all insertions consisted of two connecting phrases. For most items that received a double insertion, only one answer consisted of a double insertion. Looking at the number of double insertions per participant, we find that only a few participants inserted multiple connectives (18 of 111 participants). The data on multiple insertions therefore does not allow us to draw any strong conclusions. This will be discussed further in Section 7.5.

2.5% of all insertions in Instantiation and Specification items were manual answers (a raw count of 37 insertions). There was no clear pattern in these manual answers: only a few items received manual answers, and these items received at most two manual answers. The type of manual answer was also variable: a few of them were connectives (*although*, *while*), some were related to the syntax of the items (for example, *as of*, *in which*, and *with*), while others aimed at attributing information between the two arguments to a speaker (for example, *saying*, *stating*, *adding*). No clear conclusions can be drawn from these insertions. An additional 1.2% of the data consisted of ‘blank insertions’: subjects used the ‘manual answer’ option to not insert anything. As with the manual answers, there was no clear pattern: only a few items received a blank insertion, and there were no more than two blank insertions per item.

## 7.5 Discussion and conclusion

Instantiation and Specification are two of the most frequent types of implicit relations in the PDTB, making up 24% of all implicit relations in the PDTB. Given that PDTB and RST-DT annotators frequently disagree on the annotation of these relations, the current study aimed to investigate how readers interpret these relations. More specifically, we examined whether readers interpret them as mainly elaborative or argumentative, and we searched for characteristics that are shared by relations interpreted as argumentative ones.

The results showed that both Instantiation and Specification items received many causal insertions: 28% of insertions in Instantiation items and 39% of insertions in SPECIFICATION items were causal. Causal connectives were particularly prevalent in PDTB SPECIFICATION relations with RST-DT EVIDENCE and EXPLANATION-ARGUMENTATIVE annotations. A by-item analysis revealed that items often received two or more main types of insertions. This is consistent with a recent line of research that has focused on multiple readings of coherence relations: Rohde et al. (2015, 2016) have shown that certain relations can have more than one single reading. The current study provided more evidence for this hypothesis by showing systematic ways in which INSTANTIATION and SPECIFICATION relations can occur simultaneously.

A manual analysis of items that were predominantly interpreted as INSTANTIATION or SPECIFICATION revealed that these items often contained one of the following characteristics: (i) a larger set is mentioned in the first segment, and one member of the set is explicitly referred to in the second argument, or (ii) the first segment contained a reference to a general or vague concept. By contrast, items that were often assigned a causal label shared one of the following characteristics: (i) the first segment contains a claim, and the second segment contains evidence or an argument for this claim, or (ii) the second segment can be interpreted as a result of the situation described in the first segment.

These results go beyond previous work investigating the signaling of INSTANTIATIONS, SPECIFICATIONS and other ELABORATION relations (e.g., Li & Nenkova, 2016; Taboada & Das, 2013; Vergez-Couret & Adam, 2012). For example, Taboada & Das (2013) manually analyzed RST-DT relations, and showed that relations from the RST-DT class EXAMPLE can be signalled by individual words that indicate a relation without linking the two arguments (e.g. the word *explaining*). GENERAL-SPECIFIC relations are most often marked by entity features and lexical chains or overlap markers. Taboada & Das (2013) also found that EXPLANATION-ARGUMENTATIVE and EVIDENCE relations often do not contain any relational signal. Li & Nenkova (2016) conducted a computational corpus analysis to show that first segments in INSTANTIATION relations are often shorter than other sentences, and the second segments are often longer (compared to other types of relations). First segments of INSTANTIATIONS generally also contain fewer out-of-vocabulary words than other sentences, and more gradable adjectives (such as *high, likely*). The study reported in the current chapter differs from these efforts in that the results are based on naïve readers' interpretations of relations, rather than expert judgments.

Before turning to the implications of the results for the annotation of INSTANTIATION and SPECIFICATION relations, a critical note should be made regarding the choice of connectives. As discussed in the previous chapter, the frequency of the connectives is not controlled for in the design. The phrases *as an illustration* and *more specifically* are less frequent that *because* or *as a result.* It is possible that this played a role in participants' choices. Moreover, *as a result* might not be the most suitable connective to express an argumentative reading of the items. Other, more suitable options, might be *thus* or *therefore.* In future experiments, we aim to test whether the frequency of connectives and the semantic connotation of *as a result* influences the results by including less frequent and more subjective argumentative connectives.

Finally, the method currently does not provide clear results regarding multiple interpretations of a single item. It's possible that participants inferred two readings for a particular item but only inserted one connective. Our results are not conclusive regarding this issue, since few participants inserted multiple connectives. However, we asked participants to choose the connective that "best expresses" the meaning of the relation, in which case the chosen connective should represent the strongest reading that was inferred.

**Implications for the annotation of INSTANTIATION and SPECIFICATION items** The current chapter showed that many coherence relations annotated in the PDTB as INSTANTIATION and SPECIFICATION have a simultaneous argumentative reading. This finding supports the hypothesis that INSTANTIATIONS and SPECIFICATIONS are sometimes used to illustrate / specify a situation and to serve as an argument to a claim. PDTB focuses on the ideational relation between the arguments and therefore does not annotate the argumentative function. RST-DT does annotate INSTANTIATION and SPECIFICATION relations as elaborative or argumentative, which better converges with the dominant responses of our participants for some relations. However, neither PDTB nor RST-DT fully captures the double function of these items that was reflected in the results. In particular, RST-DT does not allow for the annotation of more than one reading per relation. PDTB annotators, while allowed to annotate more than one label per relation, hardly ever make use of this option.

The classification of INSTANTIATIONS and SPECIFICATIONS as elaborative disregards the fact that many of these relations also have an argumentative reading. However, classifying them as argumentative disregards their elaborative reading (instantiating or specifying). Consequently, neither label reflects readers' interpretations of the relation, making the annotations less cognitively plausible. In order to ensure that the annotations reflect actual reader interpretations, we recommend that both the ideational function of a relation (for example, that one segment provides an example of what is said in the other segment) and its argumentative function (for example, that the example is used to justify a claim) be annotated (also see Crible & Degand, 2017; González, 2005; Redeker, 1990).

Currently, most corpora of coherence relations do not contain multiple relation labels for a single instance. Given that readers *can* obtain two different readings for a single relation, corpora would be descriptively more adequate if relations with multiple readings would receive multiple labels. This might also improve inter-annotator agreement of these relations: when using data annotated by only two coders, differences in the annotations are interpreted as annotator error or disagreements. However, if at least one coder would annotate both readings, agreement would improve and the resulting annotations would better reflect the full meaning that the relation expresses. Of course, a double annotation process raises issues as well: it would need to be clear whether a single coder sees both senses, or whether different coders have different interpretations that are alternatives to one another but can't hold at the same time.

Alternatively, rather than annotating two separate labels for reflecting the ideational and the rhetorical function of relations, frameworks could create separate sub-classes for argumentative INSTANTIATIONS and SPECIFICATIONS. This would set relations with an argumentative function apart from purely ideational relations. This solution would be in line with the traditional assumptions that only one relation holds between two discourse segments. Adding subtypes of annotating multiple labels would likely improve the descriptive adequacy of the annotations, and thereby also the validity of the frameworks.

In sum, the current chapter discussed the double function of INSTANTIATION and SPECIFICATION relations, and showed that comprehenders can infer ideational and argumentative readings for these relations. However, we did not find evidence that readers interpret both functions simultaneously for a single item. Given that these relations can be interpreted in multiple ways, the next question that arises is whether readers show systematic biases when interpreting such multi-functional relations. In other words, do readers show individual differences when interpreting multi-functional relations as ideational or argumentative? This is what is investigated in the next chapter.

# Chapter 8

# Investigating individual differences in interpretation biases[1]

Coherence relations such as INSTANTIATION and SPECIFICATION can often convey multiple relation senses, but few studies have investigated how readers actually interpret such multi-interpretable coherence relations. The experiment reported in Chapter 7 revealed that comprehenders can infer ideational and argumentative readings for these relations. The current chapter presents the first effort to investigate whether readers have specific preferences for inferring one relational sense over another for multi-interpretable coherence relations, and whether they display individual variability in their interpretation preferences.

A repeated measures connective insertion study shows that readers have consistent biases in interpreting coherence relations, and that they differ from each other in these biases. Moreover, participants seem to differ in their "default" depth of processing, which explains some of the differences in interpretations. However, even when readers perform a task that requires them to process deeply, some differences in interpretations remain. Results from a verbal working memory task indicate that the individual differences found in the connective insertions study cannot be accounted for by differences in working memory.

There is an increasing body of evidence indicating that comprehenders display significant variation in other areas of language comprehension. The results from the current study extend these findings to the coherence relation field: readers display individual variability in their interpretation biases of multi-functional relations, which is related to how deeply they process text. Theories and experiments on discourse interpretations will therefore need to account for differences in interpretation biases as well as depth of processing.

[1] This chapter is adapted from Scholman, M.C.J., Demberg, V. and Sanders, T.J.M. (submitted). When *for example* and *specifically* can be interpreted as *because*: Individual differences in coherence relation interpretation biases. *Submitted to Discourse Processes.*

## 8.1 Introduction

In previous chapters, we have established that Instantiation and Specification relations can convey multiple senses. Example (33) illustrates this.

(33) *The two executives could hardly be more different.* **Mr. Roman comes across as a low-key executive; Mr. Phillips has a flashier personality.** wsj_1317

The second sentence of this example can be interpreted as providing examples of the differences (comprehenders can mentally insert *for example* between the sentences to mark this relation), specifying what the differences are exactly (*specifically* can be inserted to mark this relation), and/or giving an argument for the claim that these executives are different (this causal relation can be expressed by inserting *because*).[2] In the current study, we investigate whether readers systematically differ in how they interpret relations that can have multiple readings.

The different readings correspond to the two functions of the relation: the elaborative reading (i.e. giving an example or specifying) corresponds to the ideational function of relations, which involves the relation between the information conveyed in consecutive elements of a discourse (cf. Moore & Pollack, 1992). The causal reading (i.e. providing evidence to support a claim) corresponds to the argumentative function of Instantiations, which involves the writer's intention of affecting the addressee's beliefs, attitudes, etc. (cf. Hovy & Maier, 1995).

The mapping of Penn Discourse Treebank (PDTB; Prasad et al., 2008) and Rhetorical Structure Theory Discourse Treebank (RST-DT; Carlson & Marcu, 2001) relations in Chapter 4 revealed that annotators from these frameworks often interpret these relations differently: relations that are annotated as Specification or Instantiation in the PDTB are often annotated as Evidence or Explanation-argumentative in the RST-DT. These RST-DT labels are types of causal relations, where Arg2 provides evidence or an explanation for the situation presented in Arg1.

The systematic discrepancy between annotations of Instantiation and Specification relations indicates that these relations can have two different interpretations, which correspond to the ideational and the argumentative function. In Chapter 7, we conducted a connective insertion study to investigate how readers actually interpret these relations: do they interpret both functions of the relation, or do they infer one or the other interpretation? The results showed that readers do interpret certain Specification and Instantiation items as argumentative (not only ideational), but we did not find evidence that readers interpret both functions simultaneously for a single item.

Given that Specification and Instantiation relations can be interpreted in multiple ways, the next question that arises is whether readers show systematic biases

[2] The example was in fact annotated as a Specification relation in the PDTB, with CONTINGENCY.Cause.reason as a second sense.

when interpreting such multi-functional relations, and whether they differ from each other in their biases (i.e., whether there is individual variability). The current literature on individual differences in discourse interpretations does not provide an answer to these questions: little is known about how readers interpret multi-interpretable coherence relations (but see Rohde et al., 2015, 2016; Sanders, 1997), and even less is known about individual differences in the interpretations of multi-interpretable coherence relations. The current paper aims to fill this gap in the literature.

In the current study, we aim to address these issues by exploring whether highly educated comprehenders show biases in their coherence relation interpretations, and whether they differ from each other in their biases. In Study 1, we find that there are differences between participants' interpretations (some have an argumentative bias; others a non-argumentative bias) and that these differences are consistent. Participants took part in the same study with different items several weeks apart, and exhibited systematic interpretation biases. In Studies 2 and 3, we then investigate whether processing depth and working memory capacity can explain these differences in interpretations. We discuss alternative explanations and implications of these results in Section 8.6.

## 8.2 Background

There are varying streams of research investigating how readers interpret relations. Traditional coherence relation accounts presuppose that relations convey a single reading, or are associated with a single, strongest reading (e.g. Kehler, 2002; Mann & Thompson, 1988; Sanders et al., 1992, 1993). A different line of research explores the multi-functionality of relations (e.g. Blakemore, 1997; Carston, 1993; Hovy, 1995; Moore & Pollack, 1992; Sanders & Spooren, 1999). These studies all focus on the possibility of the multiple relation interpretations, but they do not address readers' actual processing and representations of the discourse. What's more, individual differences are not accounted for in this literature. The studies we review below provide the context for investigating whether readers display individual variability in their interpretations of coherence relations.

### 8.2.1 Interpretation strategies

No studies to date have examined whether readers display individual interpretation biases when it comes to coherence relations. However, results from several experiments investigating the processing and representation of coherence relations suggest that readers do have default expectations for relations that influence how they interpret a text. Two theories exist regarding possible interpretation strategies: (i) *Continuity hypothesis*: readers have a bias towards interpreting sentences and the events they describe as following one another in a continuous manner (Murray, 1997; Segal, Duchan & Scott, 1991), and (ii) *Causal preference theory*: readers have a bias towards interpreting sentences as causally related (Sanders, 2005, 2017). The second theory is of particular relevance to the current chapter, as the material that is included in our studies can be interpreted causally or non-causally.

The causal preference theory focuses on the central role that causality plays in discourse processing: studies have consistently shown that causal relations are easier to process (e.g., Black & Bern, 1981; Haberlandt & Bingham, 1978; Kuperberg, Paczynski & Ditman, 2011; Sanders & Noordman, 2000) and recalled better than non-causal relations (e.g., Myers, Shinjo & Duffy, 1987; Sanders & Noordman, 2000; Trabasso & Van Den Broek, 1985).

Some have argued that this processing advantage is surprising, since causal relations can be considered to be more complex than non-causal relations: "P causes Q" implies "P and Q" and, in general, "P precedes Q" (Noordman & Vonk, 1998). To account for this paradox, Sanders (2005, 2017) proposed the causal preference theory: when possible and in the absence of any cues indicating otherwise, readers prefer to construct a causal relation between the events. This can explain the processing difference between causal and non-causal relations: readers aim at building the most informative representation (Graesser, Singer & Trabasso, 1994; Noordman & Vonk, 1998; Sanders, 2005), so they start out assuming the relation between two consecutive sentences is a causal relation. Subsequently, causally related information will be processed faster, because the reader will only arrive at an additive relation if no causal relation can be established. Asr and Demberg (2012) and Hoek and Zufferey (2015) also find corpus-based evidence supporting this theory.

In sum, it is likely that readers have predisposed biases for interpreting relations: results from corpus-based and processing studies suggest that causal relations have a special status in discourse processing and production. However, very few studies have investigated whether individual readers differ from each other in the type of relation they infer, and whether they differ from each other in the interpretation strategies and biases that they display when processing coherence relations. This issue is addressed in the current chapter. Before we turn to Study 1, we first review literature on individual differences.

### 8.2.2 Individual differences in interpretations

Individual differences in how language is processed arise from variability in cognitive processes (e.g., working memory capacity, word decoding abilities), as well as from variability in factors such as language experience, age, interest, and motivation (Boudewyn, 2015). A large body of research has shown that individual differences affect language processing and comprehension for various areas of linguistics, such as syntax, phonology or second language acquisition (e.g., Afflerbach, 2015; Fuchs, Pape, Petrone & Perrier, 2015; Kidd, Donnelly & Christiansen, 2017).

Individual variability has also been found to be relevant in discourse processing. For example, studies have shown that less-skilled readers draw fewer inferences compared to skilled readers, are less capable of accessing background knowledge from long-term memory, and are less capable of integrating this background knowledge with the information from the text (Hannon & Daneman, 1998, 2001; Oakhill, 1984; Rapp, Broek, McMaster, Kendeou & Espin, 2007). Linderholm et al. (2000), however, found that skilled and less-skilled college readers both benefit from causal text repairs. They argue that less-skilled readers' comprehension only suffers when the task becomes more challenging because the readers will then need to recruit additional resources.

Fewer studies have looked at individual differences in the interpretation of coherence relations. Studies in this area have mainly focused on the effect of signalling on different groups of readers: skilled versus less-skilled; children of various age groups; adolescents versus adults. These studies showed that the effect of signalling varies for different types of readers, but the results vary between studies. For example, McNamara, Kintsch, Songer & Kintsch (1996) found that low knowledge readers benefit from coherence marking, whereas high knowledge readers benefit from a more implicit text. The results from a series of experiments conducted by Kamalski, Sanders and Lentz (2008) indicated that in general, low-knowledge adult readers benefit more from coherence marking than high-knowledge adult readers, but this effect is dependent on genre (linguistic marking of coherence interacts with prior knowledge in the informative genre, but not in the persuasive genre). van Silfhout, Evers-Vermeul, Mak & Sanders (2014) found that skilled adolescent readers benefit from explicit marking just as much as less-skilled adolescent readers. The difference in results between the studies can be caused by a difference in methodologies, groups of readers, and text types. These differences make it difficult to interpret and integrate the findings.

In sum, the existing literature on individual differences focuses on the specific effects of certain variables on the processing and representations of discourse by different groups of readers. The current chapter takes a different approach. We explore whether differences in processing strategies occur within a group, without taking into account specific reader characteristics such as reading skill, age, or prior knowledge. After establishing that there are indeed systematic differences between the participants in the current study, we explore a possible explanation for these differences in a follow-up study.

## 8.3 Study 1: Do readers display consistent interpretation biases?

In the current study, we investigate whether readers display different interpretation biases when they are asked to interpret ambiguous coherence relations. The participants included in this study are all highly educated: their education level ranges from an undergraduate degree to a doctorate degree.

We use the same methodology as used in Chapters 6 and 7, but for a different purpose. Rather than investigating whether certain items or relational categories receive more mixed insertions than others, we look at whether readers display individual differences in insertions. In other words, we investigate whether certain participants are more biased towards one interpretation or another.

The connective insertion paradigm was chosen because it presents new possibilities of investigating the interpretation of relations. Other methods traditionally used in discourse research (e.g., comprehension questions, recall, eyetracking-while-reading) tap into the processing ease of specific types of relations or the representation in the participants' memory. However, these methods cannot provide insight into the crucial issue that is at stake in the current studies: which relation did people infer? The connective insertion method is focused on exactly this this: it allows us to tap into the interpretations of naïve readers by providing insight into the result of the interpretation process: the relation sense that readers inferred.

The study was conducted as a repeated measures crowdsourced task: the total set of items was divided into four smaller experiments (referred to as batches) and subjects were asked to participate in each of these over the course of a few months. The results will be analyzed in terms of individual distributions of insertions per batch. This will allow us to determine whether participants differ from each other in their argumentative preference and whether the argumentative preference they display in one batch is predictive of their answers in another batch.

### 8.3.1 Participants

92 native English speakers (45 female; age range 20–68 years; mean age 39 years) participated in this study. They were recruited via Prolific and reimbursed for their participation (1.25 GBP per batch). Participants came from the United States and the United Kingdom. Their education level ranged from an undergraduate degree to a doctorate degree.

### 8.3.2 Materials

The set of experimental passages consisted of 24 implicit Instantiation and Specification relations from the aligned corpus introduced in Chapter 4 (see Appendix F). These items were also included in the connective insertion study reported in Chapter 7. The dominant response (i.e., the type of connective that was most often inserted) for these items was causal, and they can therefore be considered ambiguous between an Instantiation / Specification reading and an argumentative reading. The items were manipulated for a related study on the effect of evaluative markers, which is why there were three conditions: the original condition, a condition with additional evaluative markers in the first argument, and a condition without evaluative markers in the first argument. This manipulation will not be investigated in the current study, but it will be accounted for in the statistical analyses.

The fillers in this study consisted of 40 implicit relations: 8 Cause, 8 Conjunction, 12 Concession and 12 Contrastive relations. We chose items that were understandable without background knowledge in economics.

The experimental items were divided into three lists (because of the three evaluative markers conditions). All filler items were included in every list. Each list was divided into four batches, with 16 items per batch: 6 Instantiation or Specification items, 2 Cause, 2 Conjunction, 3 Concession, and 3 Contrast items. Order of presentation of the items per batch was randomized to prevent order effects. Subjects completed all four batches within a list over a time-period of four months, the order in which they completed the batches was randomized. Average completion time per batch was 11 minutes.

**Connecting phrases** The list of connecting phrases was identical to that used in Chapter 6 and consisted of: *as an illustration*, *more specifically*, *in addition*, *because*, *as a result*, *even though*, *nevertheless*, and *by contrast*.

### 8.3.3 Procedure

The study was distributed via Prolific and hosted on LingoTurk (Pusse, Sayeed & Demberg, 2016). The method is identical to the method reported in Chapter 6. All items were presented without linguistic context.

### 8.3.4 Results

For the analyses, we aggregated the frequencies of the connectives that fell into the same class: *because* and *as a result* were aggregated as causal connectives, and *even though* and *nevertheless* were aggregated as concessive connectives. We are mainly interested in whether participants inferred the argumentative sense of the relation. If a participant inserted multiple connectives, we therefore coded the response as causal if one of the double insertions was *because* or *as a result*.

Our analysis is similar to exhaustive, leave-one-out cross-validation: we tested the significance of every possible combination of three batches as a set for estimating bias vs. one batch for observing whether that bias was stable (i.e., 1 2 3 vs. 4; 1 2 4 vs. 3; etc.). We first recoded the participants' responses into a binary variable representing the type of responses ('argumentative' versus 'non-argumentative') and converted this to a continuous scale from 0 (no argumentative insertions) to 1 (only argumentative insertions). This variable is referred to as the prior argumentative bias. In order to be able to interpret the coefficient, we standardized the ratio by calculating the Z-score of the bias predictor variable.[3] We then modeled the results using a binomial GLMER model in the statistical software R (R Development Core Team, 2008; *lme*4 package,

[3] This was done by subtracting the mean of that variable from the original individual values and then dividing this score by the standard deviation of the variable.

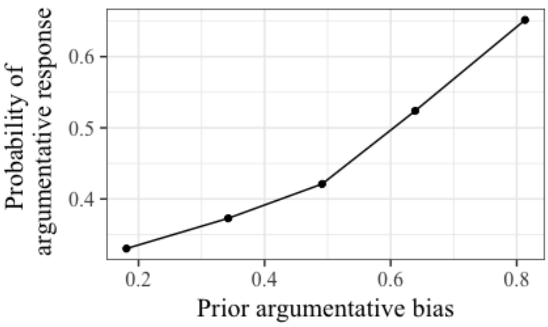


**Figure 8.1:** Probability of argumentative response as a function of the prior bias binned into five bins.

Bates & Sarkar, 2007), to determine whether a participant's prior argumentative bias displayed in responses in earlier batches is a significant predictor for responses in a new batch.

We included the evaluative markers conditions as a fixed effect in our models to explain for any variance in the insertions, but we found no significant effect of these conditions. Subject was included as a random effect, with a random slope for prior argumentative bias (the correlation was removed in order for the model to converge). A random intercept was included for item (random slope was removed for convergence).

The results show that the participants' prior bias is predictive of insertion ($\beta$ = 1.78, SE = 0.17, $z$ = 10.6, $p < .001$). This is visualized in Figure 8.1, which shows the probability of an argumentative response as a function of the prior bias.

We can interpret the coefficient as follows: for each standard deviation increase in the prior argumentative bias, the odds of a participants' response being argumentative is 5.9 times as large.

To visualize the insertions for participants with different biases, we classified participants into three groups: participants who interpreted items mainly as argumentative in at least three of four batches were classified as having an argumentative bias (22 participants); participants who interpreted items as mainly INSTANTIATION or SPECIFICATION in at least three batches were classified as having a non-argumentative bias (37 participants); and the remaining participants were classified as having no bias (33 participants). The insertions of participants in each group were grouped together, as shown in Figure 8.2. The bars represent the insertions of all participants in each group per batch; the difference of interest lies in the amount of *because / as a result* insertions (represented by the red bars on the bottom of every stack) versus *as an illustration / more specifically* insertions (light-blue and yellow bars). The results displayed in these graphs show that (at least some) participants do in fact have different interpretation biases.

Finally, 6.3% of responses for SPECIFICATION and INSTANTIATION items contained double insertions. These double insertions were mainly provided by 5 participants.

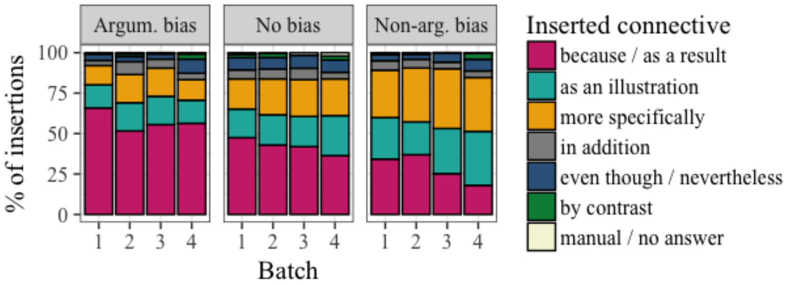


**Figure 8.2:** Distributions (%) of inserted connectives per batch for the three groups of participants.

32% of all double insertions consisted of a causal connective as well as *more specifically* or *as an illustration*.

### 8.3.5 Discussion Study 1

The results suggest that readers differ in how they interpret relations: some readers are more prone to interpret relations as argumentative, whereas others are more prone to interpret relations as non-argumentative. In other words, readers seem to differ in their argumentative bias. As to what causes this (difference in) bias, several explanations can be proposed: it could be related to individual reader characteristics, reading goals, or levels of processing. We here focus on levels of processing as a starting point in exploring the underlying cause for interpretation biases. A well-known relevant finding in earlier work on discourse processing relates to the differences between deep and shallow processing. Several studies suggest that comprehenders' reading processes are affected by the depth of processing (e.g. Aaronson & Ferres 1986; Hannon & Daneman 2004; Hayes-Roth & Walker 1979; Lehman & Schraw 2002; see also Sanford & Sturt 2002; Sanford & Graesser 2006). For example, Just & Carpenter (1978) found that readers make more inferences when instructed to perform a consistency judgment task compared to a comprehension task (which requires shallower processing). Similarly, Noordman, Vonk & Kempff (1992) found that readers who were instructed to judge the consistency of each sentence in a discourse were more likely to make causal inferences than readers that did not receive such instructions. They argue that readers are satisfied with a rather parsimonious processing in normal reading, but when they read the text with a particular goal in mind (e.g., retaining more information), they process sentences more thoroughly and engage in inference processes.

Relating the deep vs. shallow processing account to the current study, it is possible that participants with an argumentative bias might have processed the text at a different level than participants with a non-argumentative bias. An open question is whether the argumentative bias is caused by deep or by shallow processing.

A first hypothesis can be constructed based on the consistent finding in the literature that causal relations have a special status in the integration of information (as discussed in Section 8.2.1). Readers seem to have a preference for relating segments in a causal way, which results in causally-related information to be integrated faster. However, in the literature so far, existing cases of such a causal preference mechanism are limited to objective causal relations. Although the relations under investigation in the current study can be expressed with the connective *because*, they are in fact argumentative – that is, subjective causal – relations. Subjective causal relations are more complex than objective causal relations: in understanding a subjective relation, both the underlying causality (P → Q) and the epistemic proposition ("from knowing P, you may conclude Q") have to be processed (Cozijn, Noordman & Vonk, 2011). The question then is: does a causal preference also hold for subjective causal relations? If this is the case, we would expect comprehenders to interpret relations as argumentative by default – when they process text on a shallow level –, but not necessarily when they process text on a deeper level. The first hypothesis that is constructed is therefore that shallow processing is more strongly related to the argumentative bias, and thus readers with a higher argumentative bias would provide more instantiation and specification insertions when they process text more deeply.

An alternative hypothesis can be constructed based on the integration vs. inference account. Cozijn et al. (2011) show that two distinct processes occur when interpreting causal relations. The propositional integration process consists of immediately establishing a relation between the clauses. The inference process is characterized by checking the relation against world knowledge (e.g., is a causal interpretation of the relation possible, and is it true?). Studies investigating the processing of objective causal relations have shown that integration happens naturally (also during shallow processing), whereas inferences are made when processing deeply (Noordman et al., 1992; Noordman & Vonk, 1998).

Research related to the integration vs. inference account does not relate directly to the current study: most research supporting this account has investigated the generation of inferences based on objective causal relations, rather than subjective causal relations. As noted previously, subjective relations are more complex: both the underlying causality and the epistemic proposition have to be processed. However, given that subjective causal relations are so complex, and given that causal inferences are generally only made when processing deeply, it can be hypothesized that deep processing is more strongly related to the argumentative reading of the relations included in our study. The second hypothesis is therefore that *deep* processing is more strongly related to the argumentative bias, and thus readers with a non-argumentative bias would provide more causal insertions when the task requires them to process text deeper.

In order to investigate whether the shallow vs. deep processing account can provide an explanation for the individual differences in the argumentative bias, the participants will need to perform a task that forces them to process the relations more thoroughly. A summarization task does exactly this. A summary is a brief statement that represents the condensation of information accessible to a subject and reflects the gist of the discourse (Hidi & Anderson, 1986). The process of summarization is an elaborative activity which promotes deeper processing of the text; it requires participants to comprehend the individual propositions, establish connections between them, and formulate a concise and coherent verbal representation (Johnson, 1983; see also Stein & Kirby, 1992). The concise and accurate representation of the main message of the passages requires more complex integration of the information, as compared to normal reading. Summarization is therefore assumed to be an elaborative activity that facilitates deeper processing of the text.

## 8.4 Study 2: Can processing depth explain the interpretation biases?

We conducted a crowdsourcing study in which the participants from Study 1 were presented with a new set of items, divided over two batches with different tasks: the original insertion task and the summary task. We conducted both tasks to obtain a baseline of insertions for items using the original task. For the summary task, participants were first asked summarize the segments of every relation using their own words. After having provided the summary for an item, they were presented with the drag-and-drop interface and asked to insert a connective from the predefined list.

### 8.4.1 Participants

The participants from Study 1 were invited to participate in this study. 40 participants did so (19 female; age range 20–68 years; mean age 42 years). They were reimbursed for their participation (1.25 GBP per batch).

### 8.4.2 Materials

A new set of 12 implicit Instantiation and Specification relations from the PDTB were assembled for which both the ideational and argumentative reading were inferable (see Appendix F). The study also included 12 fillers, which belonged to the same relational classes as in the previous study.

### 8.4.3 Procedure

The summary task consisted of two steps for every item: first, participants were asked to summarize the item in their own words; next, they were asked to insert a connective in between the two segments.

The items were divided into two batches. Immediate linguistic context (two sentences preceding and one sentence following the segments) was included in order to facilitate the process of summarization. Participants were randomly divided into two groups: one group first completed the first batch in the original task, the other group completed the second batch in the original task. Next, these participants were invited to participate in the summary study, which contained the items of the batch they had not seen yet. The order of the items was randomised for every trial. The study was distributed via Prolific and hosted on LingoTurk.

### 8.4.4 Results

We tested whether there is a statistically significant difference between participants' insertions in the original task and the summary task, and whether a participant's prior argumentative bias as displayed in the previous study is a significant predictor for responses in a new iteration. The prior bias displayed in the previous study was stored as a continuous variable (on a scale of 0 to 1). This variable was then centered and standardized, and included as a fixed effect in the model. Task was contrast-coded and included as fixed effect. A random slope was added for task and item (correlation was removed in order for the model to converge, prior bias was not included as a random slope for convergence), as well as task and participant (correlation was removed in order for the model to converge).

We found an interaction effect between the task and the prior bias ($\beta$ = .45, SE = .23, $z$ = 2.0, $p$ < .05): the effect of the prior argumentative bias on the participants' insertions differed for the original task and the summary task. Participants with a higher argumentative bias inserted fewer causal connectives in the summary task than in the original task. Participants with a lower argumentative bias did not show a difference in their distribution of insertions between tasks.

We also found a main effect of prior bias ($\beta$ = -.52, SE = .13, $z$ = -3.99, $p$ < .001): participants with a higher argumentative bias inserted more causal connectives than those with a lower prior bias. Results also showed a main effect for task ($\beta$ = .46, SE = .21, $z$ = 2.13, $p$ < .05): fewer causal connectives were inserted in the summary task compared to the original task.

Next, we tested whether the prior argumentative bias is a significant predictor of the responses in the original task and the summary task separately. We find that, as expected, the bias is significant in the original task ($\beta$ = -.83, SE = .21, $z$ = -3.9, $p$ < .001). More interestingly, we see that the prior bias is not predictive of the response in the summary task ($\beta$ = -.26, SE = .16, $z$ = -1.59, $p$ = .1). This means that forcing readers to process more deeply diminished the effect of prior bias.

These effects are visualised in Figure 8.3, which displays the insertions into items in the original task vs. the summary task per bias type. The bias type was based on every participants' response in the previous study (i.e., participants that displayed an argumentative bias in the previous study were classified as such in the current study). Figure 8.3 shows a decrease in argumentative insertions between tasks by the

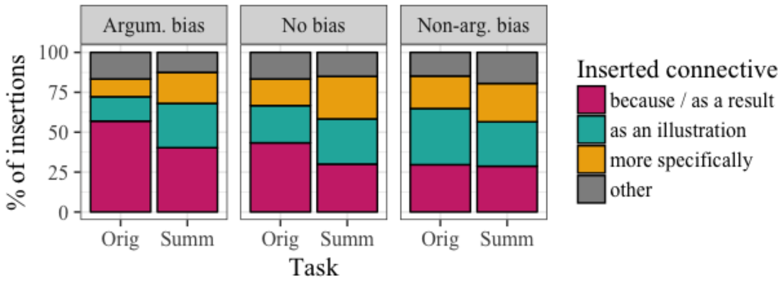


**Figure 8.3:** Distribution (%) of inserted connectives per task for the three groups of participants.

participants with an argumentative bias, also by the participants who displayed no bias, but not by participants who displayed a non-argumentative bias.

Finally, 7% of responses for experimental items in the summary task consisted of double insertions, and 1.5% of responses were manual answers. This converges with the results found in Study 1. No clear patterns could be identified.

### 8.4.5 Discussion Study 2

The biases that were identified in Study 1 were replicated in Study 2: participants who displayed a higher argumentative bias in Study 1 also interpreted more relations as argumentative in the current study, compared to those who displayed a lower argumentative bias in Study 1. However, this effect of prior bias was modulated by the task: participants with an argumentative bias interpreted fewer relations as argumentative when processing the relations deeper. Participants with a non-argumentative bias did not show a difference in their distribution of insertions between tasks. This suggests that shallow processing is related to inferring an argumentative reading, whereas deep processing is related to inferring a non-argumentative reading.

These results raise the question of why certain readers processed text deeper than others. The answer to this question could be related to possible individual differences in the readers' characteristics. Even though our participants all completed at least an undergraduate study, it can be expected that individual variability exists within this highly educated group of participants. To get more insight into the characteristics of the participants that took part in these studies, we invited them to take part in a working memory study.

Working memory (WM) is a limited-capacity short-duration system that is responsible for the temporary storage of words, phrases, and ideas, as well as the processing of new and already stored information (Baddeley & Hitch, 1974; Just & Carpenter, 1992). It is believed to be involved in a wide range of cognitive behaviors, including comprehension, reasoning, and problem-solving (Engle, 2002). Working memory capacity is also an important individual-differences variable and accounts for a large portion of variance in general intellectual ability (Conway, Cowan, Bunting, Therriault & Minkoff, 2002; Conway, Kane & Engle, 2003; Kyllonen & Christal, 1990; Süß, Oberauer, Wittmann, Wilhelm & Schulze, 2002).

Given that working memory capacity (WMC) is relevant to a large variety of cognitive skills and can explain individual differences in performance on different tasks, it seems to be a good base for investigating the source of variability in readers' biases. Several WM tasks exist, but in the current study we use the reading span test, which is known to be related to reading skill (e.g., Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Waters & Caplan, 1996). We hypothesize that the working memory component related to reading skill can explain the variability we find in readers' biases and processing depth. More specifically, people with a higher WMC possibly have more cognitive resources available to engage in deeper processing by default, which causes them to have a stronger non-argumentative bias than readers with a lower WMC. Alternatively, however, high WMC readers perhaps can afford to process text on a shallower level, given that high WMC readers generally have a higher reading skill. From this point of view, a high WMC would correlate with an argumentative bias. These hypotheses are investigated in Study 3.

## 8.5 Study 3: What is the relation between bias type and working memory?

We conducted a crowdsourcing study in which the participants from Study 2 were asked to participate in a verbal working memory test. Verbal working memory capacities were tested using an automated reading span measure (similar to Waters et al., 1987; Waters & Caplan, 1996). Participants read several sets of sentences on a computer screen and judged their acceptability. After the participants judged all sentences in the set, they were instructed to recall the last word of each sentence in the order that they read the sentences. The task thus includes both processing and storage components.

### 8.5.1 Participants

The participants from Study 2 were invited to participate in this study. 34 participants did so (18 female; age range 24–28; mean age 43 years). Participants were reimbursed for their participation (2.50 GBP).

### 8.5.2 Materials

Materials were constructed to resemble the Waters et al. (1987) reading task (see Appendix G). 56 sentences of 8 to 11 words in length were created, divided over four different sentence types: cleft subject (CS), cleft object (CO), object-subject (OS), and subject-object (SO). The sentences are illustrated in Examples (34)–(37).

(34) It was the actor that donated a large sum of money. (CS)

(35) It was the yellow notebook that the girl brought. (CO)

(36) The boy envied the friend that bought a game. (OS)

(37) The woman that the painter loved despised her parents. (SO)

These sentences vary along two dimensions: the number of propositions and the syntactic complexity. CS and CO sentences have one proposition, whereas OS and SO sentences have two propositions. CS and OS sentences are considered to be syntactically simpler than CO and SO sentences, because the thematic roles are assigned canonically in these sentences (in other words, the first noun is the agent and the second is the theme) (Waters & Caplan, 1996). As a result, the verbal working memory test contains sentences of different levels of complexity, with CS sentences (one proposition and syntactically simple) being the easiest and SO sentences (two propositions and syntactically more complex) being the most difficult in the test.

Half of the sentences of each type contain a verb that requires an animate subject (e.g. *donate*, *forget*) and half contain a verb that requires an animate object (e.g. *fascinate*, *disappoint*). Additionally, half of the sentences is acceptable and half is unacceptable. Unacceptable sentences were formed by inverting the animacy of the subject and object noun phrases (e.g. *It was the girl that the yellow notebook brought*).

### 8.5.3 Procedure and analysis

We developed an automated version of the reading span test following recommendations by Unsworth, Heitz, Schrock & Engle (2005) and von der Malsburg & Vasishth (2013) for automated operation span tasks, and Conway et al. (2005) for working memory span tasks in general. The study was distributed via Prolific and hosted on LingoTurk.

The test consisted of three phases. In the first practice phase, participants had to judge the acceptability of 8 sentences. The reaction time was measured for each judgment. The average reaction time plus two standard deviations was used as a cutoff point (i.e. time-out) in later stages (cf. von der Malsburg & Vasishth, 2013; Unsworth et al., 2005), to prevent participants from taking extra time to rehearse the words, and from writing down the word elsewhere. This individual time-out time allows participants to work at their own pace.

In a second practice phase, participants did the full task: they judged the acceptability of sentences and memorized the final word of every sentence. At recall, participants were instructed to write the words from the current set in the correct order. Participants were instructed to perform the acceptability task very accurately and then perform as best as they could on the recall task (cf. Waters & Caplan, 1996). This phase consisted of two blocks – one of 3 sentences and one of 5 sentences – to allow participants to get used to blocks of different sizes.

In the main test, set sizes from 2 to 5 sentences were presented (as recommended by Conway et al., 2005); there were four sets of each size (cf. von der Malsburg & Vasishth, 2013). The sets were presented in a random order so that participants could not anticipate how many words they had to remember before they were instructed to recall them. The total study took 15 minutes on average.

Working memory capacity was calculated using the partial-credit unit (PCU) scoring procedure (as recommended by Conway et al., 2005; Friedman & Miyake, 2005). PCU expresses the mean proportion of words within a set that were recalled correctly, and ranges from 0 to 1. Typos were accounted for by allowing for a one-character difference between the target word and the response.

### 8.5.4 Results

The working memory capacity of our group of participants ranged between .29–.99, with an average of .74 (SD = .19). This range compares relatively well to the range of working memory reported in Friedman & Miyake (2005), although it is slightly more extreme: the working memory capacity of students participating in their study ranged from .46–.9, with a mean of .68. The increased variability in the range of PCU can be explained by the difference in educational level (undergraduate up to doctorate degree) and age (mean age 43 years) in our population compared to theirs (undergraduate students, mean age not reported). Indeed, the participants who scored lower than .46 (the minimal PCU found in Friedman & Miyake, 2005) in our study were all older than 45 years. We therefore conclude that the distribution of working memory found in this study is relatively prototypical of participants with our demographics.

Results showed that there was only a marginally significant correlation between working memory capacity and prior bias (as displayed in Study 1) (Pearson's $r$ = -0.31, $p$ = .07). In order to determine whether PCU is a significant predictor of response, the GLM model from Study 2 was repeated with PCU as a predictor variable. The model included main effects for PCU, task and prior bias, as well as the interactions for PCU:task and task:prior bias. Random intercepts for item and participant were also included (random slopes were removed for convergence). Results showed no interaction effect between task and PCU ($\beta$ = -.62, SE = 1.31, $z$ = -.47, $p$ = .63), nor a main effect of PCU ($\beta$ = -.16, SE = .79, $z$ = -.2, $p$ = .84). The interaction between task and prior bias remained significant ($\beta$ = .51, SE = .26, $z$ = 2, $p < .05$), as well as the main effect of prior bias ($\beta$ = -.54, SE = .15, $z$ = -3.6, $p < .001$).

### 8.5.5 Discussion Study 3

Given that working memory capacity is involved in comprehension and reasoning, it was expected that differences in working memory would be related to discourse interpretations as well. However, the results showed no significant relation between working memory and participants' bias, nor an effect on participant's response (to either of the insertion tasks). This suggests that working memory capacity is not related to the type of discourse interpretations or interpretation biases that are relevant in our experiments.

Of course, it is possible that the sample size was too small to detect any effects. The PCU was relatively high overall (with a mean of .74), and so there were too few participants with a lower working memory capacity. However, the skewed distribution is a common characteristic of PCU, and this has not been an issue in other studies. Another possible explanation for the absence of an effect is the type of task used to measure working memory. Reading span is correlated with comprehension. Perhaps a different task, which taps into a different component of working memory, would be more related to discourse interpretations. However, interpreting discourse requires comprehending it, and so it seems likely that the reading span task is the most suitable task for investigating the relation with discourse interpretations. Based on the results of this study, we therefore conclude that WMC cannot explain the variability in interpretation biases and processing depth that we find in Study 1 and 2.

## 8.6 General discussion and conclusion

Coherence relations can often be interpreted in different ways or convey multiple relation senses (see Rohde et al., 2016; Sanders, 1997; Sanders & Spooren, 1999; Webber, 2013). Few studies have investigated how readers actually interpret multi-interpretable coherence relations: do readers have a common systematic bias to interpret such relations in a certain manner? Or do they show individual differences in their interpretation preferences? The results from this study suggest that readers differ in how they interpret relations, at least when it comes to INSTANTIATION and SPECIFICATION relations with multiple functions: some readers are more prone to interpret relations as argumentative, whereas others are more prone to interpret them as ideational. The results from Study 2 revealed that these biases interact with the depth of processing. The responses of readers with a non-argumentative bias do not change when they perform a task that requires deep processing, whereas the responses of readers with an argumentative bias become less argumentative when they process text more deeply.

The results from Study 2 highlight the importance of the task: readers interpret text differently depending on the task they are asked to perform. This issue has perhaps not received enough attention in experimental discourse studies. Magliano & Graesser (1991) and Graesser, Singer & Trabasso (1994) discuss the problems this causes. Specifically, experiments might not measure general interpretations and reading processes because the task requires only shallow processing. If the experimental task does not require deeper processing, readers might adopt an artificial reading strategy that prevents them from making certain interpretations or inferences. It is therefore important that researchers reflect on what level of processing their task requires or stimulates.

Study 3 showed that the bias and the depth of processing do not depend on differences in working memory capacity. Future research should focus on identifying causes for this individual variability. One area of interest could be background knowledge. Graesser et al. (1994) argue that readers employ shallower processing if they lack the background knowledge that permits the establishment of specific inferences. Background knowledge (or familiarity with the topic) can therefore be considered a prerequisite for the generation of online inferences (Cozijn et al., 2011; Noordman et al., 1992). Building on this, Noordman & Vonk (1998) contend that the nature of the reading process depends on the reader, and, in particular, on the reader's knowledge. Relating this to the current study, one could argue that perhaps the participants with a higher argumentative bias had less background knowledge to rely on, in which case they reverted to shallower processing. The level of background knowledge of our participants in relation to our material is difficult to measure, since we used short excerpts from different news articles in the eighties and nineties. It was assumed, however, that all participants had a similar (low) level of background knowledge with regards to the topics discussed in the materials. The relation of background knowledge to the interpretation of coherence relations with multiple readings presents an interesting avenue for future research.

There are a couple of limitations to the experimental design chosen in our experiments. First, the absence of context in Study 1 could have had an influence on the results. It is generally assumed that readers successively construct a conception of the writer's goals while reading a discourse. They do this based on the content that they have encountered in the discourse so far. Using a paraphrase study to investigate the interpretation of coherence relations in context, Sanders (1997) found that readers are more likely to interpret ambiguous relations as subjective when the linguistic context is also subjective (but see Canestrelli et al., 2016). In the current study, items were presented without the context, and so participants were able to construct different conceptions of what the context could have been like, which in turn could influence their decisions. However, it should be noted that participants were in fact explicitly informed about the original genre of the text. This should have helped in constructing a conception of the writer's goals (given that they were aware that the writers were Wall Street Journal journalists). Moreover, no effect of context was found in the connective insertion study reported in Chapter 6. Finally, participants were presented with context in Study 2, and the results of the original condition replicated the bias found Study 1.

The finding that readers differ from each other in their biases for inferring an argumentative or ideational relation contributes to theoretical accounts of these two functions. As noted in Chapter 2, many accounts of coherence agree on the existence of separate functions of relations (the *multi-level thesis*, cf. Moore & Pollack, 1992, see also Sanders & Spooren, 1999), but they disagree on how the functions relate to each other. For example, Moore & Pollack (1992) assume that, in discourse interpretation, the recognition of one function follows from the recognition of the other function. Whether the ideational or argumentative function is recognized first depends on the relation, the context, and the reader's knowledge. Mann & Thompson (1988) and Hovy (1995) assume that, for every relation, one function will be prevalent or primary, and so one label will be more suitable for that relation. None of these accounts have considered readers' actual processing and representation of the discourse, as in experimental studies. The approach used in the current study does take into account readers' interpretations of relations, and shows that what is considered the more prevalent reading is not the same across individuals: readers systematically differ in their biases for inferring one reading over the other.

The possibility of systematic differences in readers' interpretations of coherence relations has been largely overlooked in the field. Individual variability has been an annoyance rather than a relevant factor to researchers, and as a result, individual variation has often been considered as "error variance" in experiments (see also Kidd, Donnelly & Christiansen, 2017). Studies that did investigate differences between readers have focused on group differences rather than individual differences. However, there is an increasing body of evidence indicating that comprehenders display significant variation in other areas of language comprehension. The results from the current study extend these findings to the coherence relation field: readers that can be characterized as belonging to the same group (i.e., educated participants) display individual variability in their interpretation biases of multi-functional relations, which is related to how deeply they process text.

The current paper presented the first exploration of this issue, and will hopefully prompt further investigations into individual variability in discourse comprehension, and in the processing of multi-interpretable relations. Several directions of future research can be identified. First, regarding multi-functional relations, an interesting avenue is to investigate readers' interpretation biases using other types of relations that frequently correlate (e.g., Motivation and Condition, as discussed in Moore & Pollack 1992). Second, as previously discussed, the nature of the biases deserves more investigation: what causes readers to infer one reading over another? Third, the relation between these results and online processing should be explored. The connective insertion task is an off-line, meta-linguistic method. Perhaps readers with a non-argumentative bias employed a deeper processing in the task than they would in a less meta-linguistic task (such as eyetracking-while-reading). In other words, one can wonder whether the non-argumentative interpretation is forced by the task or whether it is something that occurs naturally. The question then is, would they still have a bias for inferring the non-argumentative function in a more naturalistic task? It should be noted that this is difficult to test: on-line studies target the processing of discourse, but provide no information about the resulting representation of that discourse.

Finally, the effects of the biases on online processing provide an interesting area for future research. We know that objective causal relations are easier to process than non-causal objective relations (cf. Canestrelli et al., 2013; Traxler et al., 1997). It can be asked whether readers with an argumentative bias (interpreting relations as subjective causal) process multi-functional relations faster than readers with a non-argumentative bias (inferring a non-causal reading). This would therefore test the boundaries of the causal preference theory (Sanders, 2005). Concerning the processing of multi-interpretable relations in general, we can ask whether processing of such relations requires more effort, compared to the processing of less ambiguous relations. Additionally, the question of whether readers are able to infer multiple readings deserves more attention. If some readers are in fact able to infer multiple readings of a single relation during online processing, do these readers take longer to process the relation, given that they hold multiple readings in their mental discourse representation?

Based on the results of the studies presented in this chapter, we conclude that readers have consistent biases in interpreting coherence relations, and that they differ from each other in these biases. Moreover, participants seem to differ in their "default" depth of processing, which explains some of the variability in interpretations. However, even when readers perform a task that requires them to process deeply, differences in biases occur. Theories explaining and experiments investigating discourse interpretations will therefore need to account for differences in depth of processing as well as in interpretation biases.

# Chapter 9

# Investigating the construction of discourse structure during comprehension[1]

Connectives provide processing instructions regarding the way in which two discourse segments should be related to each other. The results from the PDTB–RST-DT mapping reported in Chapter 4 indicate that comprehenders (or at least annotators) rely heavily on connectives, rather than on the content of the segments, to infer a relation: in absence of an explicit connective, annotators show relatively little agreement on the type of relation they infer. The current chapter investigates whether readers make use of both the information provided by the connective and the content of the segments during discourse processing.

The cue phrase *On the one hand* (OT1H) is used to test whether the expectation for contrast that is set up by this cue phrase can be satisfied by any contrastive connective following OT1H, or whether readers are able to take into account the content of the segments and establish a contrastive relation with the appropriate segment – even if that might not be the first contrastive segment following OT1H. We used short passages containing *On the one hand* and *On the other hand* (OTOH) as well as an intervening contrastive sentence marked by *but*. There were two versions of this intervening contrastive sentence: a locally contrastive sentence did not establish an appropriate contrast with the OT1H sentence; a globally contrastive intervening sentence did. If comprehenders take into account the content of the segments, they should be sensitive to the difference between locally and globally contrastive sentence.

Three experiments using offline and online measures show that comprehenders anticipate more than just a matching form (i.e., *On the other hand*); rather, they are

[1] This chapter is adapted from, and in some parts identical with, Scholman, M.C.J., Rohde, H. and Demberg, V. (2017). "On the one hand" as a cue to anticipate upcoming discourse structure. *Journal of Memory and Language*, 97, 47–60.

able to construct complex discourse structures and establish non-local dependencies. The results reveal that comprehenders maintain their expectation for an upcoming contrast across intervening material, doing so even if the embedded constituent itself contains contrast. Furthermore, comprehenders disprefer a subsequent contrast marked with *On the other hand* when a passage has already provided intervening content that establishes an appropriate contrast with *On the one hand.*

## 9.1 Introduction

Connectives and cue phrases (from now on referred to as 'connectives') provide processing instructions regarding the way in which two discourse segments should be related to each other. There is ample evidence that readers make use of connectives to link segments together (e.g., Britton, 1994; Canestrelli, Mak & Sanders, 2013; Gernsbacher, 1997; Murray, 1997; van Silfhout, Evers-Vermeul & Sanders, 2015). More recent investigations provide evidence that comprehenders can also use these cues to make predictions about which coherence relations to establish within or between sentences (Köhne & Demberg, 2013; Rohde & Horton, 2014; Xiang & Kuperberg, 2015).

Another area of research that has provided support for the importance of connectives is discourse annotation: annotators can agree on the relational sense of explicit relations relatively well. This can be attributed to the information that the connectives provide regarding the type of relation that holds. The mapping study between PDTB and RST-DT reported in Chapter 4 also revealed that the agreement between annotators working in different frameworks is acceptable for explicit relations. However, this same study showed that agreement on implicit relations was relatively low, indicating that inferring a relational sense in the absence of a connective might be more difficult. This raises the question of how sensitive readers actually are to the content of the segments in the presence of a connective. The studies reported in this chapter investigate this sensitivity: do comprehenders apply a "lazy" strategy after encountering a clear signal for a relation by paying less attention to the content of the segments, or do they still process the segments deeply enough to be sensitive to the attachment height of segments in the discourse structure?

A starting place to address this issue is the presence of connectives that constrain possible upcoming discourse structures. Consider (38).

(38) John is considering a job at the Edinburgh zoo.
On the one hand, he really needs the money, because he should start paying off his student loans.
On the other hand, he hates the idea of cleaning out panda cages.

The first sentence in (38) introduces an issue that is up for debate (i.e. accepting a job offer). The next two sentences present two contrasting perspectives (contrast1 and contrast2), marked with *On the one hand* (OT1H) and *On the other hand* (OTOH), respectively. The marker OT1H signals to the comprehender to expect an upcoming contrast2, an expectation that typically must be satisfied for the passage to be considered an acceptable discourse.

Although OT1H is strongly constraining, it does not fully determine the nature of the next sentence. For example, contrast2 can be signaled by a marker other than OTOH, such as *but* or *however*. Likewise, contrast1 itself does not necessarily need to be marked with OT1H, as evidenced by the acceptability of (38) without the explicit OT1H marker. Furthermore, the expectation for contrast2 does not need to be satisfied immediately; intervening material can occur between contrast1 and contrast2, as is the case with the *Also* sentence in (39).

(39) John is considering a job at the Edinburgh zoo.
On the one hand, he really needs the money, because he should start paying off his student loans.
Also, his car needs to be serviced.
On the other hand, he hates the idea of cleaning out panda cages.

To understand the passage in (39), comprehenders must infer a contrast relation between the OT1H and OTOH sentences even though they are not adjacent, and establish that the *Also* sentence is part of the constituent conveying the contrast1 perspective. An intervening sentence can also be part of a contrastive relation without specifically contrasting with the OT1H contrast, as illustrated by the intervening sentence in (40), which contrasts with the previously embedded material on the topic of student loans.

(40) John is considering a job at the Edinburgh zoo.
On the one hand, he really needs the money, because he should start paying off his student loans.
But the loans could be deferred for a few more months.
On the other hand, he hates the idea of cleaning out panda cages.

In this case, comprehenders must infer that the intervening sentence provides a contrast with the *because*-clause, and therefore does not satisfy the expectation for a contrast set up by OT1H. If readers do not take into account the content of the intervening sentence after reading OT1H, they would likely interpret the OTOH sentence as presenting contrast3. As such, passages with OT1H/OTOH allow us to investigate whether readers consider both the connective and the content of the segments when they establish coherence relations. Moreover, the pair of markers allows us to test the generation and maintenance of discourse-level expectations across multiple sentences. This extends prior work on the processing of discourse markers (e.g., Canestrelli, Mak & Sanders, 2013; Drenhaus, Demberg, Köhne & Delogu, 2014; Xiang & Kuperberg, 2015) and coherence relations (e.g., Köhne & Demberg, 2013; Rohde & Horton, 2014; Mak & Sanders, 2013) by moving beyond relationships that hold within sentences or between adjacent sentences, and instead investigating how specific markers can guide structure building on a large scale during processing.

The studies presented in this chapter test whether readers use both the information provided by a discourse marker and the content of the segments to process a coherence relation across intervening material. One possibility is that comprehenders take any contrastive connective to satisfy the prediction of contrast2 based on OT1H, in which case an OTOH sentence after the intervening contrastive sentence should yield an infelicitous discourse structure (that is, a three-sister structure that violates the constraint imposed by OT1H for a binary contrastive relation). However, if comprehenders build and maintain rich discourse structures (Kuperberg, 2016), they will link contrast2 to contrast1 even if additional content intervenes that conveys another type of contrast.

The chapter is laid out as follows. The next section reviews related work on cue-based anticipation. Section 9.3 then describes the design of the materials used in our experiments. The rest of the chapter presents three studies. Study 1 is a story acceptability study, and functions as a norming study for our materials. Study 2 is a story continuation study, designed to test comprehenders' anticipation regarding the content of the next sentence. Study 3, an eye-tracking while reading experiment, tests whether structure-sensitive biases emerge during online processing.

## 9.2 Background

Several studies have investigated whether comprehenders use cues to facilitate the processing of coherence relations, but little is known about how the presence of connectives interacts with the information that is available in the segments. The studies we review below provide the context for investigating whether comprehenders take into account both the connective and the content of the segment when processing coherence relations.

### 9.2.1 The effect of connectives on discourse processing

Existing work points to the role connectives play in guiding comprehenders' processing of discourse. For example, Canestrelli, Mak & Sanders (2013) investigated whether the information that connectives provide is used during the online processing of causal relations. They focused on the Dutch connectives *want* (a prototypical marker of subjective causal relations) and *omdat* (a prototypical marker of objective causal relations), which are both translated as *because* in English. In three eye-tracking studies, they demonstrated that the subjective connective *want* leads to an immediate processing disadvantage compared to *omdat*. This effect was observed at the words immediately following the connective, at which point readers cannot yet establish the causal relation on the basis of the content. The effect hence seems to be solely induced by the connectives, indicating that readers immediately adjust their interpretation of relations based on the connective, without taking into account the content of the segments.

Canestrelli et al. (2013) investigated the effect of connectives marking subtle differences within the class of causal relations. Other studies have shown that comprehenders can also immediately update their expectations of relations based on connectives that come from very different relational classes. For example, Drenhaus et al. (2014) conducted an ERP study to investigate the processing of relations signaled by the concessive connective *however* or the causal connective *therefore.* They found that the connective yielded an immediate effect: late positivities were elicited at *however* in comparison with *therefore.* This is reflective of an updating process: readers immediately exploit the information that the connective provides about the upcoming relations, and they adjust their expectations accordingly (switching their anticipation of a causal relation to an until then unexpected concessive relation) (see also Köhne & Demberg, 2013; Xiang & Kuperberg, 2015; Xu, Jiang & Zhou, 2015). These results emphasise the importance of connectives during discourse processing, and the speed with which they are processed.

Prior work has also shown that paired markers can cue the relation before the first segment is even presented. Staub & Clifton (2006) found that comprehenders use the markers *either/or* to predict a coordination structure within sentences. In (41), the content after *or* can attach at the sentence-level or at the noun phrase (e.g., *Either Linda bought the red car or the green one.*).

(41) (Either) Linda bought the red car or her husband leased the green one.

The results from an eye-tracking study revealed that participants read the *or*-clause faster when *either* was present. This indicates that *either* cued the participants to expect an upcoming coordination. In addition, the results indicated that readers misanalyzed the sentence coordination as noun-phrase coordination when *either* was absent, which led to longer reading times at the end of the sentence coordination condition (compared to a noun-phrase-only coordination condition: *The workers painted (either) the house or the barn over the summer*). Readers were therefore able to predict an upcoming sentence coordination based on the word *either*, which facilitated processing of that structure once they encountered it.

The demonstrated effects just described are for connectives that express the relation at hand. *On the one hand* elicits an expectation for a subsequent sentence that will express a relevant contrast, but this expectation does not have to be satisfied immediately and can be expressed by different connectives (e.g., *but* or *on the other hand*). The new studies we report here therefore target what effect a "misleading" connective has on the interpretation of discourse. We present readers with passages containing OT1H, which sets up an expectation for contrast2, and which is then followed by an intervening contrastive sentence marked by *but.* The question is whether readers are able to infer from the content if the intervening sentence presents contrast2, or whether they rely solely on the connective and are therefore not sensitive to content of the intervening sentence. Section 9.3 provides further details on the design of the studies. First, we consider the occurrence of OT1H and OTOH in natural text.

### 9.2.2 Occurrence of OT1H and OTOH in natural text

In order to get more insight into the nature of OT1H and OTOH in language, we first look at the occurrence of these markers in natural text. All fragments containing one or both markers were extracted from the ukWaC corpus, a 2 billion-word corpus of English UK webpages (Baroni, Bernardini, Ferraresi & Zanchetta, 2009). For every occurrence of OT1H, the following three paragraphs were searched for OTOH. Similarly, for every occurrence of OTOH, the preceding three paragraphs were searched for OT1H. This amounted to 60,749 instances containing one or both markers. The passages revealed several features of OT1H and OTOH that show why they are suitable for the current experimental aims.

First, we find that the appearance of OT1H or OTOH does not wholly determine the presence of the other: only 18% of passages contain both OT1H and OTOH. In 3% of the data OT1H occurs without OTOH (from an OT1H perspective, this means that 17% of all occurrences of OT1H are not followed by OTOH). In such cases, OTOH was replaced by other connectives and cue phrases (most commonly *but*, *at the same time* and *while*), which suggests that natural text often requires readers to link contrast2 to contrast1 even without the OT1H/OTOH pairing.

In 7% of the OT1H/OTOH-marked data, other sentences intervene between the OT1H-sentence and the OTOH-sentence. Importantly, this intervening material can itself contain coherence relations (see Appendix H for an example of this). The intervening material sometimes also contained contrastive markers. For example, *but* occurred between OT1H and OTOH pair in more than 500 passages. *But* is therefore an ambiguous marker when it follows OT1H: it can mark contrast2, but also other contrastive relations that are presumably embedded within contrast1. Such contexts require readers to process an embedded coherence relation, while also maintaining an expectation of contrast2.

These distributions hence attest to the frequent presence in naturally occurring contexts of the types of complex discourse structures that will be targeted in this chapter.

## 9.3 Experimental design and predictions

The experimental items consist of contexts like (42), in which the type of additional material intervening before OTOH is varied, as well as the presence/absence of OTOH-marked contrast2.

(42) **Intro:** Joseph got a job offer from the Edinburgh Zoo and he's pondering whether he should take it.
**OT1H:** *On the one hand*, he needs the money that this job will pay,
**Cause:** *because* he should start paying off his student loans this year.

(i) **Global contrast:** But he could keep looking for a nicer, better-paying job.

(ii) **Local contrast:** But the loans could be deferred for a few more months.

(iii) **No contrast:** Also, his car needs to be serviced by the end of the month.

**OTOH:** *On the other hand*, he hates the idea of cleaning out panda cages and lion dens every day.

The first sentence in (42) introduces a situation in which a decision regarding a certain action is considered. The second sentence presents contrast1, which is an argument in favor of one particular decision. This sentence has a subordinate clause marked by *because*, which presents an explanation for the decision. One of three different types of intervening sentences can follow this OT1H-sentence. The global contrast condition is shown in (i) with a sentence marked by *but*. The content of the global contrast plausibly contrasts with the content of the OT1H-sentence and satisfies the contrast1∼contrast2 pairing (take the job vs. keep looking for a job). The local contrast condition is shown in (ii) with a sentence marked by *but*. The content contrasts with the information embedded in the subordinating *because*-clause that directly precedes it (having to pay off loans vs. deferring loans) and does not fully satisfy the contrast1∼contrast2 pairing. The baseline condition is shown in (iii) with a sentence marked by *also*, whose content does not contrast with any preceding information.

Figure 9.1 illustrates the attachment heights of the intervening sentence in the local and global contrast conditions. This distinction can be tested by omitting the because-clause: if the intervening sentence is still felicitous in the story, the contrast that this intervening sentence provides is global rather than local. To illustrate this, consider Example (42-ii) again, but without the *because*-clause: "On the one hand, he needs the money that this job will pay. But the loans could be deferred for a few more months." In this scenario, the loans mentioned in the local contrast are not introduced and the intervening sentence is infelicitous. In the global condition, the subordinate clause can be omitted without the intervening sentence becoming infelicitous, as illustrated by Example (42-i): "On the one hand, he needs the money that this job will pay. But he could keep looking for a nicer, better-paying job."

Passages such as (42) allow us to test whether comprehenders can make use of both the information provided by the connectives (e.g., OT1H, *also*/*but*, OTOH) and the content of the segments (including the intervening sentence). If comprehenders take into account both the connectives and the content of the segments for the construction of coherence relations and are sensitive to these structural distinctions between the

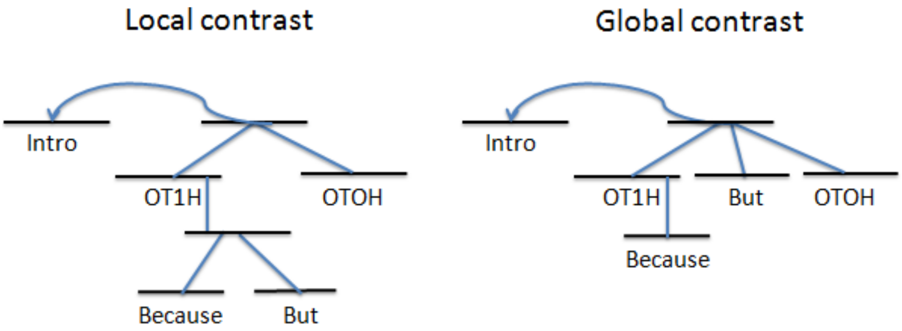


**Figure 9.1:** Attachment heights of the *But*-sentence in the local and global contrast conditions.

intervening sentences, the OTOH-sentence is predicted to be disfavored in the global condition, because it presents an unnecessary or unexpected "third hand" compared to the baseline and local conditions. However, if comprehenders do not rely on the content of the segments and therefore do not consider the attachment height of the intervening sentence, the results will show no difference between the global and local condition.

## 9.4 Experiment 1: Norming study

The current studies rely on a manipulation of the additional intervening sentence (see (42), where the global contrast condition discusses better-paying jobs, the local contrast condition discusses loans, and the no-contrast condition discusses a car). Crucially, we want to avoid a scenario where any observed effects are attributable to basic properties of the information in those intervening sentences, rather than our manipulation itself. If participants find one story variant more entertaining or topically more interesting (e.g., jobs vs. loans?), their behavior in the subsequent experiments might reflect those judgments rather than the discourse parsing that is studied in Experiments 2 and 3.

We therefore collected naturalness ratings for a superset of stories, selecting for the later experiments a subset for which participants gave similar ratings to the two contrast conditions, under the assumption that they found the content of both similarly interesting and natural. The experiment had a 3 × 2 design: type of intervening sentence (global/local/no contrast) × presence/absence of the final OTOH-sentence. The presence of the OTOH-sentence was varied in order to determine the naturalness of the item at the precritical region in the later studies, right before a participant would encounter (or write a continuation for) the final sentence. The OTOH-absent condition mirrors the materials for the story continuation task; the OTOH-present condition corresponds to the eyetracking task.

### 9.4.1 Participants

144 native English speakers (age range 18-75 years; mean age 35 years; 77 female), registered as ‘workers’ on the Mechanical Turk website hosted by Amazon, received monetary compensation for their participation ($1 per batch). Participants had various educational levels ranging from high school to a Master’s degree.

### 9.4.2 Materials and Procedure

Thirty-one items were created for this study, with the intention of selecting the 24 most suitable items. All had the same structure as (42). The presence or absence of the final OTOH-sentence was varied, creating six conditions. The 24 experimental items can be found in Appendix I.

Each participant rated the naturalness of one version of 10 or 11 stories (because of the uneven number of items), and 8 filler items on a scale of 1-7 (7 as most natural). Every version of each item was rated by eight people. Filler items consisted of short stories that were either unnatural or natural, to create a spectrum of naturalness for assessing participants’ understanding of the rating scale. Unnatural filler stories contained discourse violations such as an incorrect connective given the context (e.g., ‘so’ where it should be ‘because’) or an unexpected turn of events (e.g., receiving a bonus after having lost an important judicial case).

### 9.4.3 Analysis Methods

Results from all experiments reported in this chapter were modeled using linear mixed-effect regression models (LMER; Baayen, Davidson & Bates 2008). Models were evaluated using $lme4$ package within the statistical software R (Bates & Sarkar, 2007; R Development Core Team, 2008). For binary response variables, binomial mixed effects regression models were used. The models were always constructed with maximal random effect structure. Random effects were only reduced in cases of non-convergence of the full model (Barr, Levy, Scheepers & Tily, 2013). In these cases, the model was simplified by first removing correlation between random intercepts and random slopes (lmer model: Y $\sim$ X + (1+X || item) + (1+X || subject)), and then removing random intercepts or random slopes (lmer model: Y $\sim$ X + (1 | item) + (1 | subject)). All cases where maximal models did not converge are reported together with the results of the maximal converging model.

The factor for OTOH presence/absence was centered, and the 3-level factor for intervening sentence was deviation coded. Significance of fixed effects was evaluated by performing likelihood ratio tests, in which the fit of a model containing the fixed effect for each condition is compared to another model without it but that is otherwise identical in random effects structure. All pairwise comparisons were conducted using subsets of the data, only including the observations for the relevant pairs of

| | **OTOH-absent** | | **OTOH-present** | |
|---|---|---|---|---|
| | M | SD | M | SD |
| Global contrast | 4.75 | 1.71 | 4.78 | 1.63 |
| Local contrast | 4.65 | 1.74 | 4.99 | 1.60 |
| No contrast | 4.23 | 1.89 | 5.65 | 1.47 |

**Table 9.1:** Mean rating (and standard deviation) of the naturalness of the items per condition.

conditions with re-centered predictors. For the fixed factors, the regression coefficient, the standard error, the *t*-value, and the corresponding *p*-value associated with the likelihood ratio test are reported. For the fixed factors with more than two levels, only the *p*-value and the degrees of freedom are reported, as the likelihood ratio test only evaluates the difference between models with/without that entire factor.

### 9.4.4 Results

To obtain a set of 24 similarly acceptable items for the remaining studies, the seven stories with the lowest rating in any of the conditions were excluded from this analysis and from the subsequent experiments. The mean ratings for the conditions, based on the 24 included items, are given in Table 9.1. Rating scores of the set of 24 items were modeled using linear mixed-effect regression models, as described above.

The ratings show several patterns. First, as per the goals of the norming study, there is no main effect of intervening sentence type ($p$ = 0.42). There is, however, a main effect of OTOH presence, whereby the OTOH-present condition was rated as more natural than the OTOH-absent condition ($\beta$ = 0.584, SE = 0.16, t = 3.64, $p <$ 0.001), but this is driven by an interaction between OTOH presence and intervening sentence type (multi-level factor so only model comparison p-value reported: $p <$ 0.001). To understand the nature of the interaction, the OTOH-absent and OTOH-present conditions are considered in turn.

Looking at only the OTOH-absent stories, the results show that some contrast is better than having none, and the lowest ratings are for the no-contrast condition (a main effect of intervening sentence type ($p <$ 0.05, 2 d.f.)). Pairwise comparisons reveal a difference between ratings in the global condition and the non-contrastive condition ($\beta$ = 0.54, SE = 0.18, t = 3.09, $p <$ 0.01) and between the local condition and the non-contrastive condition ($\beta$ = 0.39, SE = 0.18, t = 2.24, $p <$ 0.05). Crucially, however, the two contrast types (global vs. local) do not differ significantly from each other ($\beta$ = -0.12, SE = 0.17, t = -0.74, $p$ = 0.46).

The OTOH-present stories provide the flip side of this: stories with no intervening contrast received the highest ratings and again the two contrast types did not differ from each other. The likelihood ratio test showed a main effect of intervening sentence type in OTOH-present stories ($p < 0.001$, 2 d.f.). Pairwise comparisons revealed a difference between ratings in the global and non-contrastive conditions ($\beta$ = -0.87, SE = 0.22, t = -3.90, $p < 0.001$) and between the local and non-contrastive conditions ($\beta$ = -0.652, SE = 0.15, t = -4.26, $p < 0.001$). Again, no significant difference was found between the global and local conditions ($\beta$ = 0.223, SE = 0.23, t = 0.98, $p$ = 0.32).

### 9.4.5 Discussion Experiment 1

This story acceptability study served as a norming study for the following experiments. The goal was to ensure that any effect that might be found in the remainder of this chapter between the global and the local condition is not due to a difference in naturalness of the stories. The results show no difference between the global and local condition, which indicated that the acceptability of the conditions will not be a confounding factor in the other experiments. The results also show that non-contrastive intervening sentences between OT1H/OTOH resulted in more natural-sounding stories than either locally or globally contrastive intervening sentences. This is not surprising, as OT1H and OTOH are typically used to express a relation with two contrastive situations, rather than three. Section 5 presents the story continuation study using the OTOH-absent materials and Section 6 presents the eye-tracking study using the OTOH-present materials.

## 9.5 Experiment 2: Story continuation study

Participants were asked to write a story continuation to the version of the items without the OTOH-sentence. The goals of this study were twofold: first, it tests whether readers interpret both a globally and a locally contrastive sentence marked by *but* as satisfying the expectation for a contrast set up by OT1H; second, the current study gives us insight into which connectives participants prefer to use to signal contrast with the OT1H clause.

### 9.5.1 Participants

90 native English speakers (age range 18-61 years; mean age 33 years; 54 female), registered as ‘participants’ on the Prolific Academic website received monetary compensation for their participation (1.50 GBP per batch). Participants had various educational levels ranging from high school to a Master’s degree.

### 9.5.2 Materials and Procedure

The experimental stimuli consisted of the 24 stories that were selected based on acceptability judgments, see Section 9.4. The stories did not contain the OTOH sentence, and appeared in three conditions: the sentence following the OT1H sentence contained a global contrast, a local contrast, or no contrast. Each version of every item was completed by ten people. Each participant saw one version of 8 stories, and 10 fillers items. Filler items consisted of short stories in the same format as the experimental items, without the markers OT1H and OTOH. Participants were asked to read the sentences and then write two sentences to continue the story.

### 9.5.3 Annotation

The continuations were manually coded for the following properties:

- Explicit marking: whether a connective was used, and if so, which one.
- Coherence relation: whether the content of the continuation contrasted with the OT1H-clause or not.

For determining the coherence relation, the two sentences that participants were asked to write were considered; specifically, it was determined whether one of the provided sentences presented a contrast with the content in the OT1H-clause. To help determine this, a connective insertion test was used with the intervening sentence excluded: if a contrastive connective could be inserted directly between the OT1H-clause and the provided continuation, the continuation was marked as contrastive with OT1H. Passage 43 shows an example of the insertion test for a continuation in the local contrast condition (intervening sentence in brackets).

(43) *Prompt:* Frank is thinking about quitting his job at the supermarket after working there for five years. On the one hand, he thinks he could get a more promising job at a multinational, because he studied accounting in college. [But he has no real work experience as an accountant.]
*Continuation:* (implicit 'On the other hand') Perhaps he could intern or do an apprenticeship. There are so many other options to consider.

### 9.5.4 Results

For the analysis of the coherence relations expressed in the continuations, the binary outcome of continuation type was modeled using mixed-effect logistic regression models, with likelihood ratio tests to compare models differing in the presence or absence of the fixed factor for condition. Models included random intercepts and random slopes.

Figure 9.2 shows the distribution of completions by condition. The global condition yielded the fewest continuations that contrasted with the OT1H sentence (10%, when [+contrastive +OTOH] continuations are collapsed with [+contrastive –OTOH continuations]); the local condition yielded more (35%), and the no-contrast condition

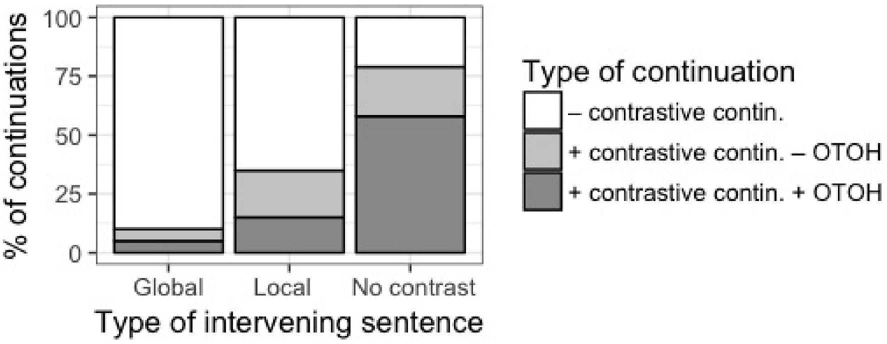


**Figure 9.2:** Types of continuations, by type of intervening sentence.

yielded the most (79%). A likelihood ratio test confirmed a main effect of condition ($p < 0.001$, 2 d.f.; for convergence, the correlations for the random slope of item were removed, as well as the random slope for subject). Pairwise comparisons confirmed that the global condition differed from the local condition ($\beta$ = 2.69, SE = 0.87, z = 3.1, $p < 0.001$), which in turn differed from the no-contrast condition ($\beta$ = -3.66, SE = 0.81, z = -4.51, $p < 0.001$). This means that participants were sensitive to the global / local contrast manipulation, and that especially the globally contrastive sentence strongly reduced the need for a subsequent contrast.

In 57.8% of all continuations, participants included an overt connective, providing insight into how often OT1H is followed by the explicit OTOH expression and other connectives. The most common connective in continuations that contrasted with OT1H was OTOH (61.3% of +contrastive continuations). The most common non-OTOH markers used to signal +contrastive continuations were *however*, *but*, and *although*. Appendix J shows the full breakdown of connective use across continuation types and conditions.

### 9.5.5 Discussion Experiment 2

The results from the story continuation study showed that participants distinguished contrastive from non-contrastive intervening material (global/local vs. no-contrast) and, moreover, distinguished different types of intervening contrast (global vs. local). These results suggest that comprehenders do anticipate a specific discourse structure after encountering OT1H and take into account the content of the segments and their attachment height in order to construct their formulation of a following utterance. Furthermore, although most of the contrastive markers that participants provided were OTOH, other contrastive connectives were also used. This shows that although OTOH is the preferred marker after OT1H, other markers can also signal contrast2.

The current study provided evidence that comprehenders are able to take into account the information expressed in the segments and attach this at the appropriate place in the discourse structure. Moreover, it provided insight into the types of connectives that people use to complete stories with OT1H. However, it does not show whether comprehenders are able to take into account both the connective and the content of the segments during real-time discourse processing. In other words, are readers able to immediately attach incoming (intervening) sentences to previously read content and take into account both the information provided by the connective and the content of the segments? These questions are investigated in Experiment 3.

## 9.6 Experiment 3: Eye-tracking study

An eye-tracking study was conducted to test whether readers are able to take into account both the connective and the content of the segments during online discourse processing. The set of experimental items consisted of the same 24 items used in the off-line studies.

An additional condition was included to investigate whether OT1H leads readers to predict an upcoming OTOH. For this condition, referred to as the OT1H absent condition, OTOH occurred without being preceded by the OT1H marker, and the intervening sentence was non-contrastive (marked by the connective *also*). The reading times of OTOH in the OT1H absent condition will be compared to those in the no contrast condition. If readers immediately build expectations of discourse structures based on cue phrases, we would expect that readers will construct an expectation of an upcoming OTOH phrase after encountering OT1H. Consequently, the presence of OT1H should lead to faster processing of OTOH.

### 9.6.1 Participants

39 native speakers of English participated in this experiment, 7 of which had to be excluded (4 due to eye-tracking problems and 3 due to problems with the computer). Participants were recruited from the University of Edinburgh community. Data from the remaining 32 participants (age range 18-30 years; mean age 22 years; 18 female) was analyzed. All participants had normal or corrected-to-normal vision. Participants were paid for their participation (15 GBP for 90 min) and were unaware of the purpose of the experiment.

### 9.6.2 Materials

The experimental stimuli consisted of the same 24 items used in the off-line studies. These items occurred in three conditions: three of these varied in their intervening sentence (global, local, and no contrast), and the fourth contained the no contrast intervening sentence without the explicit marker OT1H. This OT1H absent condition also ensured that OTOH was not always preceded by an OT1H cue in the experiment.

48 stories for two unrelated experiments were also included as fillers, along with 12 filler items containing aspects of all three experiments. In these filler items, OT1H

sometimes occurred without OTOH, or the other way around. This was done to approximate the rate at which OT1H and OTOH occur together in natural text. Note that even if participants come to expect OTOH after having seen OT1H, this can only serve to decrease our chances of seeing a difference across the global, local, and no-contrast conditions.

The stimuli were counterbalanced across four lists, with each story appearing in a different condition in each list. All participants saw each story in only one condition. The participants were randomly assigned to one of the lists, and for each participant, the list was presented in a unique, random order.

### 9.6.3 Procedure

Participants were tested individually. They were seated at a distance of approximately 60 centimeters from the monitor. Participants' eye movements were recorded with SR Research Eyelink 1000 at the sampling rate of 500 Hz. Viewing was binocular, but only the participant's dominant eye, as determined by a parallax test prior to the experiment, was analyzed. Participants rested their head on a chin-rest. Their movements were not restricted, but they were instructed to move as little as possible during the eye-tracking part of the experiment.

Each session started with an oral instruction, after which the eye-tracker was adjusted if necessary. A brief calibration procedure was then performed. This procedure was repeated after a short break halfway through the experiment and whenever measurement accuracy appeared insufficient. Upon successful calibration, the experiment started with three practice trials. The participant was instructed to read the passage at a natural pace and press the space bar after reading the entire story. Before presentation, a fixation mark appeared, first in the middle of the screen and then at the position of the first word of the first sentence. The stories were presented randomly and in their entirety on the screen. It was ensured that the critical region "On the other hand" never occurred at the beginning or end of a line. To encourage participants to read carefully, each story was followed by a comprehension question. The questions were answered using the 'F' key for *no* and the 'J' key for *yes*. The eye-tracking component lasted approximately 45 minutes.

### 9.6.4 Analysis procedure

For analysis purposes the sentences were divided into four regions, as illustrated in (44):

(44) (But he could keep looking for a nicer,) / better-paying job. $_{pre-criticalregion}$ / On the other hand, $_{criticalregion}$ / he hates $_{spillover1}$ / the idea $_{spillover2}$ / (of cleaning out panda cages and lion's dens every day.)

The critical region was the OTOH region, as this is where the reader could have difficulties due to misanalysis depending on how the preceding sentence is aligned in the discourse structure. The two words preceding this region were the pre-critical region.[2] The first spillover region contained the two words following the expression OTOH and the second spillover region contained the third and fourth word following OTOH.

Three reading time measures were computed: first pass duration, regression path duration and total reading time. First pass duration is the time spent in a region before moving on or looking back. This measure reflects the immediate processing difficulties a reader has when reading a region for the first time (Rayner, 1998). Regression path duration is the summed fixation duration from when the current region is first fixated until the eyes enter the next region on the right. This measure thus includes regressions to regions to the left of the current region. Regression path duration can be seen as reflecting the process of integrating the linguistic material with the previous context (Rayner, 1998). Total reading time is the total time spent in a region, including regressions to that region.

Prior to all analyses, skipped regions were treated as missing data. Additionally, fixations shorter than 80 milliseconds and longer than 2000 milliseconds were removed. It is assumed that the reader did not process any linguistic input during fixations shorter than 80 milliseconds, and that fixations longer than 2000 milliseconds reflect tracker losses or indicate that the participant was distracted. In all reading time measures, outliers were removed by excluding reading times more than three standard deviations from both the participant's mean and the condition's mean in a region (1.7% of all data).

### 9.6.5 Results

Reading times were modeled using linear mixed-effect regression models, with subjects and items as crossed random effects. As in the previous experiments, likelihood ratio tests were computed to compare mixed-effects models differing only in the presence or absence of the fixed factor for condition. We first discuss the results for the comparison between the global, local and no contrast condition. Then we will discuss the results for the comparison between the no contrast and OT1H-absent condition (both containing the *Also* intervening sentence). Table 9.2 shows the mean reading time measures per condition and region; Figure 9.3 shows the total fixation duration per type of intervening sentence for the global, local and no contrast condition.

No difference in reading time was found at the pre-critical region in the first pass duration ($p = 0.17$, 2 d.f.), regression path duration ($p = 0.26$, 2 d.f.), and total fixation duration ($p = 0.38$, 2 d.f.).

[2] The two words preceding the pre-critical region were not analyzed, but were included in Figure 9.3 for illustrative purposes.

| | **Region** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Pre-critical* | | *Critical* | | *Spillover 1* | | *Spillover 2* | |
| | M | SE | M | SE | M | SE | M | SE |
| *First pass durations* | | | | | | | | |
| Global contrast | 282 | 132 | 352 | 146 | 276 | 129 | 231 | 116 |
| Local contrast | 240 | 114 | 318 | 134 | 261 | 124 | 237 | 109 |
| No contrast | 259 | 139 | 322 | 131 | 242 | 106 | 232 | 106 |
| OT1H absent | 246 | 114 | 324 | 127 | 258 | 117 | 223 | 111 |
| *Regression path durations* | | | | | | | | |
| Global contrast | 409 | 228 | 391 | 190 | 306 | 173 | 288 | 216 |
| Local contrast | 382 | 255 | 369 | 227 | 297 | 215 | 298 | 235 |
| No contrast | 364 | 241 | 351 | 188 | 274 | 173 | 289 | 203 |
| OT1H absent | 338 | 217 | 367 | 192 | 278 | 169 | 255 | 157 |
| *Total reading time durations* | | | | | | | | |
| Global contrast | 335 | 148 | 453 | 221 | 326 | 170 | 272 | 148 |
| Local contrast | 311 | 152 | 400 | 178 | 309 | 176 | 285 | 160 |
| No contrast | 303 | 161 | 387 | 165 | 293 | 146 | 297 | 173 |
| OT1H absent | 283 | 129 | 401 | 196 | 291 | 146 | 256 | 129 |

**Table 9.2:** Mean reading times and standard deviations in milliseconds per measure and region.

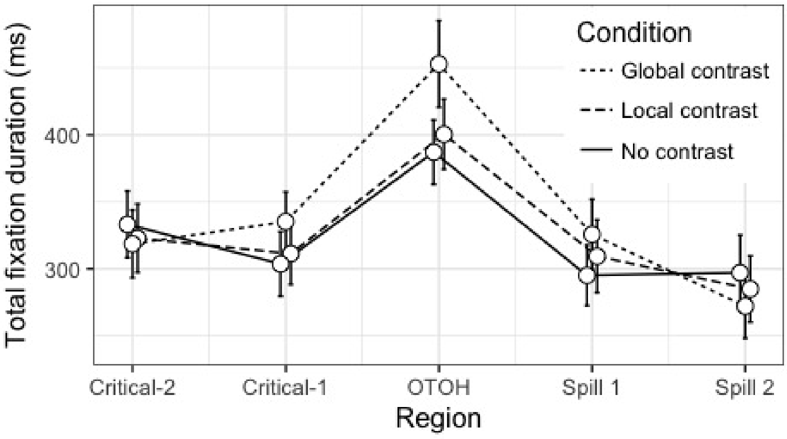


**Figure 9.3:** Total fixation duration by experimental region, per type of intervening sentence.

| | **Pairwise comparison** | | | |
|---|---|---|---|---|
| | *Global – no-contrast* | | | |
| | $\beta$ | SE | t | $p$ |
| First pass | 30.72 | 14.18 | 2.17 | $<.05^1$ |
| Total fixation | 67.68 | 24.15 | 2.80 | $<.01$ |
| | *Global – local* | | | |
| First pass | -35.06 | 13.45 | -2.61 | $<.05$ |
| Total fixation | -50.86 | 20.50 | -2.48 | $<.05^1$ |
| | *Local – no-contrast* | | | |
| First pass | -3.29 | 14.83 | -0.22 | .82 |
| Total fixation | 16.93 | 20.23 | 0.84 | .40 |

**Table 9.3:** Regression coefficients and test statistics from linear mixed-effects models for the effect of condition on the reading times of the critical region. [1]Model did not contain correlation between random slope and random intercept of item.

At the critical region *On the other hand*, the results showed a main effect of intervening sentence type in the first pass duration ($p < 0.05$, 2 d.f.) and total fixation duration ($p < 0.05$, 2 d.f.). Pairwise comparisons were conducted to see whether the three conditions differed significantly from each other. These are presented per reading time measure and condition in Table 9.3. No significant difference in regression path duration was found at the critical region ($p = 0.19$, 2 d.f.).

For the first pass duration, the results revealed a significant difference between the global and no-contrast conditions, and between the global and local conditions: reading times of OTOH were longer in the global condition than in the local and no-contrast conditions. No significant difference was found in first pass duration of OTOH between the local condition and the no-contrast condition.

For the total fixation duration, a similar picture emerges: the global condition differs from the no-contrast condition, as well as from the local condition: the total fixation duration was longer in stories with the globally contrastive sentence than in stories with the locally contrastive or non-contrastive sentence. No significant difference was found between the no-contrast and local conditions.

For spillover region 1, no effect of intervening sentence type was found in the first pass duration ($p = .11$, 2 d.f.), regression path duration ($p = .43$, 2 d.f.) or total reading time measure ($p = .28$, 2 d.f.). Similarly, no effect was found at spillover region 2 for first pass duration ($p = .81$, 2 d.f.), regression path duration ($p = .92$, 2 d.f.) or total reading time ($p = .37$, 2 d.f.).

Even though the reading times at the precritical region did not differ significantly between conditions, Figure 9.3 does show a slight difference in reading times at the

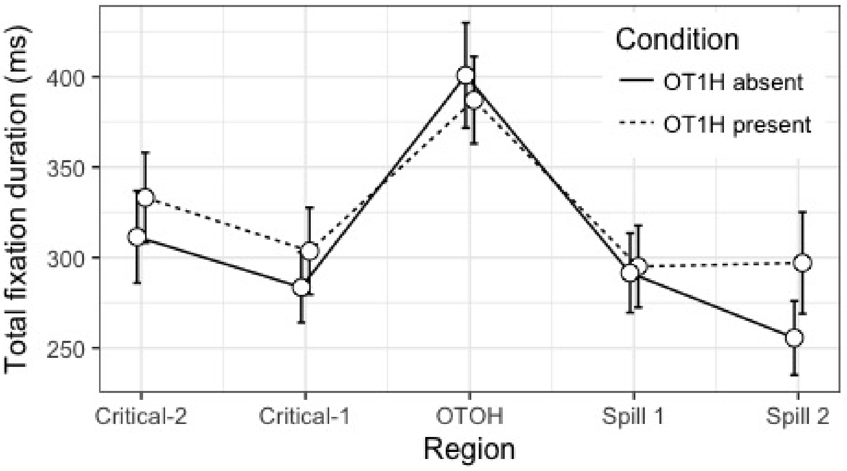


**Figure 9.4:** Total fixation duration by experimental region, for the OT1H-present (no contrast) condition and OT1H-absent conditions.

precritical region. To ensure that any effect found at the critical OTOH region is not caused by a spillover effect from any (non-significant) difference at the precritical region, we ran new models that include the reading times at the precritical region as a predictor.

For the first pass duration, the results still showed a main effect of intervening sentence type ($p$ = 0.05, 2 d.f.). Pairwise comparisons again showed a significant difference between the global and local conditions ($\beta$ = -25.7, SE = 11.42, t = -2.25, $p < 0.05$; the model did not include a correlation between the random intercept and random slope of item). The difference in reading times in the global and no-contrast conditions were only marginally significant when precritical reading times were included as a predictor ($\beta$ = 21.18, SE = 10.97, t = -1.84, $p$ = 0.06). The difference between the local and no-contrast conditions remained non-significant ($\beta$ = -5.55, SE = 11.19, t = -0.49, $p$ = 0.62). For the total fixation duration, the previously significant effect became marginally significant when precritical reading times were included as a predictor ($p$ = 0.06, 2 d.f.). This will be addressed in the discussion.

Finally, we compared the reading times of OTOH in the no contrast condition and the OT1H-absent condition. Figure 9.4 shows the total fixation duration for these conditions. The results fail to show a facilitative effect of OT1H on the processing of OTOH in any of the regions for any of the measures (pre-critical region: first pass duration ($\beta$ = 21.52, SE = 14.43, t = 1.49, $p$ = .14), regression path duration ($\beta$ = 22.78, SE = 28.38, t = 0.80, $p$ = .42), total reading time ($\beta$ = 21.52, SE = 14.43, t = 1.49, $p$ = .14); critical region: first pass duration ($\beta$ = -2.50, SE = 12.92, t = -0.19, $p$ = .84), regression path duration ($\beta$ = -15.59, SE = 19.44, t = -0.80, $p$ = .42), total reading time ($\beta$ = -13.66, SE = 21.02, t = -0.65, $p$ = .51); spillover region 1: first pass duration ($\beta$ = -16.73, SE = 11.54, t = -1.45, $p$ = 15), regression path duration

($\beta$ = -2.94, SE = 24.07, t = -0.12, $p$ = .90), total reading time ($\beta$ = 3.69, SE = 15.86, t = 0.23, $p$ = 0.81); spillover region 2: first pass duration ($\beta$ = 7.31, SE = 12.10, t = 0.61, $p$ = .54), regression path duration ($\beta$ = 33.81, SE = 27.22, t = 1.24, $p$ = .21), total reading time ($\beta$ = 41.43, SE = 20.68, t = 1.54, $p$ = 0.13)).

### 9.6.6 Discussion Experiment 3

The results showed that participants' reading times of OTOH were longer in the global contrast condition than in the local condition. This demonstrates comprehenders' ability to build discourse structures that distinguish between superficially similar intervening constituents. Even though both the local and global intervening sentences start with *but*, the underlying structures are different; only the global condition satisfies the expectation for a relevant contrast, and participants show sensitivity to this difference.

The results indicate some spillover effect from the precritical region on the reading times of the critical OTOH region: the difference between the global and no-contrast conditions was significant in the original model of first pass duration at the critical region, but marginal in a model that included precritical reading times; the main effect of intervening sentence type on total fixation duration was likewise marginal in the larger model. These findings could be attributed to differences in the final words of the intervening sentences (which differ across conditions), although this would require that the differences are systematic enough to differentiate the conditions and influence effects at the critical region. A review of the materials reveals no immediately evident bias in the construction of the intervening sentence.

Alternatively, two other factors could be at play – parafoveal processing and the reduced power for models with an additional factor. Parafoveal processing (Rayner, 1998) is known to occur when readers process words surrounding the current fixation (the parafovea extends out to 5 degrees on either side of fixation). Considering that OTOH is a marked phrase (after being pre-activated by OT1H), it is likely easily identifiable from parafoveal vision. A follow-up experiment could address this explanation by postponing the presentation of OTOH: if another intervening sentence occurs before OTOH and is identical across conditions, any increase in reading times for the global condition at this new intervening sentence would provide evidence that the increase is caused by parafoveal processing. In addition to parafoveal processing, the inclusion of the precritical reading times as an extra factor in our post-hoc analyses may have also served to reduce the models' power.

Although effects are commonly observed in spillover regions, our results were limited to the critical region itself. This likely reflects the nature of the critical region: "On the other hand" is a marked and quite long phrase, possibly already pre-activated after encountering "On the one hand". The length of the region may have provided sufficient time for participants to resolve any difficulty in integrating OTOH after encountering an intervening contrast.

Regarding the exploration of an effect of the presence of OT1H on the processing of OTOH, the results failed to show any facilitative effect. Given that the results did show that comprehenders create discourse expectations across larger fragments, we had in fact expected to find evidence for a facilitative effect of OT1H on processing of OTOH. However, the conclusion that OT1H does not influence the processing of OTOH should be viewed with some care, since it is based on the absence of a main effect. One possible explanation for the absence of an effect is that the facilitative effect of OT1H was reduced because the eye-tracking experiment contained a large number of items with OTOH (36 items out of 84 items total), which increased the expectedness of OTOH overall. Another possible explanation is that comprehenders are used to encountering OTOH in natural language without first encountering OT1H (given that 79% of all occurrences of OTOH in the ukWaC corpus were not preceded by OT1H). Readers therefore might not experience a lot of processing difficulty upon encountering an unannounced OTOH. This issue will be investigated further in the next chapter.

The results from the eye-tracking experiment can be summarized as follows. First, reading times of OTOH were longer when the expression was preceded by a globally contrastive sentence than when it was preceded by a non-contrastive sentence or a locally contrastive sentence. Crucially, the difference between the global and local conditions remained when taking into account the reading times of the precritical region. No significant difference in reading times of OTOH was found between the no-contrast condition and the local contrast condition. These results indicate that the expectation for a contrast set up by OT1H is satisfied by a globally contrastive sentence, but not by a locally contrastive sentence. This means that readers are able to take into account the information provided by the connective as well as the content of the segments during processing. When encountering a globally contrastive sentence, their prediction for a contrast is satisfied. This then leads to processing difficulty when encountering OTOH, which is reflected in additional reading times.

## 9.7 General discussion and conclusion

The current chapter addresses the question of whether comprehenders take into account the content of the discourse segments during processing of texts containing an explicit, “misleading” connective. More specifically, we investigated using offline and online tasks whether readers are able to properly attach an intervening contrastive sentence in the discourse structure of texts containing the set of OT1H/OTOH markers.

The results of the story continuation study showed that passages with an OT1H-marked contrast1 and a locally contrastive sentence were continued with sentences that convey contrast2 significantly more than passages with a globally contrastive sentence. This indicates that participants were indeed sensitive to the content of contrastive sentences that specify their local or global attachment height. Although most of the markers that participants provided were OTOH or a compound thereof, other contrastive connectives such as *but* and *although* were also used. This suggests that the anticipation of contrast2 was not merely a surface expectation for the OTOH expression.

The eye-tracking study tested whether readers build structure-sensitive representations of the discourse during on-line processing. The results showed that OTOH was processed more easily following a non-contrastive or locally contrastive intervening sentence, compared to a globally contrastive intervening sentence. This indicates that readers take into account the content of the segments in the presence of a ("misleading") connective, whereby only a contrastive sentence that targets contrast1 specifically can satisfy the expectation for contrast2.

What can be concluded, then, about the processing of coherence relations? The current study has shown that comprehenders do not rely solely on the available connectives when interpreting a discourse. The fact that the expression *On the other hand* is dispreferred more when it follows a globally contrastive sentence than a locally contrastive sentence supports the hypothesis that readers are able to create specific expectations of discourse structure based on the content of the constituents. Moreover, the fact that they are able to maintain such predictions across sentences can be taken as evidence that readers build predicted discourse structures immediately, rather than wait until the end of a sentence to integrate the full discourse structure. Readers are then able to update this expected structure as soon as they encounter evidence that their current representation is false, as evidenced by the longer reading times at OTOH.

The current studies also address a persistent gap in discourse research: while there has been a wealth of research on the processing relations within or between adjacent sentences, there has been a lack of work on processing of coherence relations whose segments are presented across larger spans of text (with intervening material). The present work makes an important contribution, because it identifies and tests a type of cue that constrains upcoming discourse in a way that can be analyzed and tested. The studies on OT1H/OTOH hence open up a new domain by taking a first step towards an understanding of what comprehenders' strategies are for building up a representation of a whole discourse while incrementally perceiving new input. An open question regarding predictive discourse processing is why the presence of OT1H did not facilitate the processing of OTOH. This issue will be investigated further in the next chapter.

# Chapter 10

# Comparing the facilitative effect of "On the one hand" in English and Dutch

Prior research has shown that people anticipate upcoming linguistic content, but most work focused on phonological, syntactic, and lexical-semantic predictions rather than predictions of coherence relations. The results from the eye-tracking study reported in Chapter 9 provide insight into whether readers are able to maintain expectations of discourse structure across larger spans of text. Readers were presented with passages in various conditions. In one condition, passages contained a contrast relation marked by "On the one hand" (OT1H) and "On the other hand" (OTOH), and an intervening sentence between them. In another condition, OT1H was absent, thereby leaving the first segment unmarked. Contrary to predictions, no facilitative effect of the presence of OT1H was found on processing of OTOH.

The current chapter explores two hypotheses explaining this null effect. First, we hypothesized that a possible facilitative effect of OT1H was reduced because the expectedness of OTOH overall increased throughout the study, due to a high number of OTOH items. We therefore conducted an English eye-tracking study with a lower proportion of OTOH items. The results revealed a facilitative effect of OT1H on OTOH, which suggests that the frequency of OTOH in the earlier study did indeed conceal a possible facilitative effect.

A second hypothesis was that OT1H does not have a strong facilitative effect in English because OTOH appears often in natural language without the prior foreshadowing of OT1H. Given that OTOH is preceded by OT1H more frequently in Dutch than in English, it was investigated whether the facilitative effect of OT1H is greater in Dutch. We conducted a Dutch eye-tracking study, similar to the English one, which revealed a stronger facilitative effect of OT1H in Dutch compared to English. This suggests that the effect of discourse markers on comprehenders' processing is modulated by language-specific factors such as distributional characteristics, and can therefore differ cross-linguistically.

## 10.1 Introduction

There is a growing body of research suggesting that readers and listeners make predictions about upcoming sounds, words and syntactic structures based on information that they have encountered so far (see Kuperberg & Jaeger, 2016 for a review). However, this prior work has focused largely on anticipation of relationships within the sentence (e.g. Altmann & Kamide, 1999; Arai & Keller, 2013; Clifton, Frazier & Connine, 1984; DeLong, Urbach & Kutas, 2005; Federmeier & Kutas, 1999; Staub & Clifton, 2006; Yoshida, Dickey & Sturt, 2013) or locally between adjacent sentences (Arnold, 1998; Van Berkum, Brown, Zwitserlood, Kooijman & Hagoort, 2005; Kehler, Kertz, Rohde & Elman, 2008). Across multi-sentence passages, the possible relationships to be established are much more flexible than those afforded by sentence-internal phonological, syntactic, and lexical-semantic constraints. Given this flexibility, an open question is whether and how comprehenders manage expectations regarding cross-sentence relationships.

The current chapter is focused on expectations of coherence relations. Such relations are assumed to hold both locally between adjacent sentences and remotely across intervening clauses (as shown empirically in corpus annotations, e.g., RST-DT: Carlson et al., 2003; PDTB: Prasad et al., 2008; Discourse Graphbank: Wolf & Gibson, 2005). Very few constraints dictate the nature of the growing discourse structure such as where or how subsequent utterances will attach (cf. Polanyi, 1988). This raises the question of whether comprehenders anticipate how upcoming coherence relations will link future sentences to the prior discourse, and whether they do so across multi-sentence passages in which the growing structure may be very open-ended.

The few studies that have investigated whether readers make predictions regarding coherence relations specifically have all focused on relations within sentences or between adjacent sentences (e.g., Drenhaus et al., 2014; Köhne & Demberg, 2013; Rohde & Horton, 2014; Xiang & Kuperberg, 2015). The previous chapter presented several studies whose results can be interpreted as extending these findings: the studies indicate that readers make fine-grained predictions of upcoming discourse structure across multiple sentences. Specifically, when readers encounter the marker "On the one hand" (OT1H), they anticipate an upcoming contrast relation and build a corresponding discourse structure. Readers maintain this structure during processing, and are able to embed intervening constituents before they encounter the second segment of the contrast relation, marked by "On the other hand" (OTOH). However, when an intervening sentence marked by *but* has provided a plausible contrast with the OT1H sentence, the reading times of OTOH increase, indicating processing difficulty. This is interpreted as evidence that readers build and maintain predictions of upcoming contrast based on OT1H.

Surprisingly, however, results from the eye-tracking study did not reveal a facilitative effect of the presence of OT1H on processing of OTOH. Several possible explanations can be found for the absence of an effect. First, it should be noted that the study contained a large number of items with OTOH: 36 items out of 84 items in total. It is possible that this repeated exposure to items with OTOH led participants to adapt to this marker, and thereby also adapt their expectations. Evidence for rapid expectation adaptation was found in a self-paced reading experiment conducted by Fine, Jaeger, Farmer & Qian (2013), who showed that repeated exposure to a priori unexpected structures can reduce, and even completely undo, their processing disadvantage. Relating these results to the eye-tracking study reported in Chapter 9, it can be hypothesized that the facilitative effect of OT1H could have been reduced because the expectedness of OTOH overall increased throughout the study.

A second possible explanation for the absence of an effect of OT1H on the processing of OTOH is related to the distribution of OT1H and OTOH in natural language. As discussed in Chapter 9.2.2, OTOH appears often in natural language without being preceded by OT1H (79% of all occurrences of OTOH in the ukWaC corpus), but OT1H is followed by OTOH relatively often (17% of all occurrences of OT1H in the ukWaC corpus). Given that it is so natural for OTOH to occur without OT1H, it can be hypothesized that OTOH without OT1H is not difficult to process. This hypothesis relates to the notion of surprisal (Hale, 2001; Levy, 2008): the difficulty of a word is proportional to its surprisal (its negative log-probability) in the context within which it appears. A word's surprisal is relatively low when that word is likely to appear in a given context (i.e., in the current study: the phrase OTOH when OT1H appears in the context), and is high when a word is less likely. A word's surprisal is thus also a measure of its expectancy (Levy, 2008).

Since OTOH occurs frequently without being foreshadowed by OT1H, the processing difficulty (caused by surprisal) of OTOH without OT1H might not be very high. This, however, also depends on the frequency of OT1H without OTOH: if OT1H is frequently followed by a marker other then OTOH, the expectancy of OTOH in the context of OT1H will be lower compared to when OT1H is always followed by OTOH. In order to be able to estimate the effect of these distributional characteristics, we need to compare the facilitative effect of OTOH in a language that contains a similar marker but with different distributions.

The current chapter presents a series of experiments that follow up on these hypotheses. To test the first explanation, we conducted an eye-tracking study in which participants were presented with a smaller number of items containing OTOH, thereby reducing a possible adaptation effect throughout the study. To test the second explanation, we conducted a similar eye-tracking study in Dutch, containing the Dutch equivalents of OT1H/OTOH ("aan de ene kant"/"aan de andere kant"). The comparison between the processing of OTOH in English and Dutch can provide insight into the issue of whether OT1H does not have a strong facilitative effect in English.

Dutch was chosen because the English and Dutch pairs behave similarly, but the distributions differ slightly between languages: approximately 63% of all occurrences of Dutch OTOH in the SoNaR corpus (Oostdijk, Reynaert, Hoste & Schuurman, 2013) are not foreshadowed by Dutch OT1H (compared to 79% in English). Moreover, OT1H is followed by OTOH relatively frequently in both languages: 17% of all occurrences of OT1H in the ukWaC corpus is followed by OTOH vs. 25% of all occurrences of Dutch OT1H in the SoNaR corpus. In other words, OTOH can occur without OT1H in both languages, but does so more frequently in English, and OT1H frequently occurs with OTOH in both languages (note that the expectations of OTOH evoked by English and Dutch OT1H are tested in a story continuation study, here presented as Experiment 1). If these distributional differences are strong enough to affect readers' expectations, processing of OTOH in passages without OT1H should be easier in English than in Dutch, and hence the presence of English OT1H provides less opportunity for a facilitative effect.

## 10.2 Experimental design and predictions

The experimental items were adapted from the items used in Chapter 9. Each item occurred in two versions: one in which OT1H was present, and one in which OT1H was absent. The items were interspersed with other passages for unrelated studies in both eye-tracking experiments. In order to ensure that the items within these studies were comparable, several changes were made to the experimental passages. Below, the items used in the English and Dutch studies are described in more detail, including the adaptations. It should be noted that these adaptations can also affect reading times of OTOH, which is a possible confound of this experiment. We will elaborate on this issue in the discussion.

### 10.2.1 English materials

The experimental stimuli consisted of 16 three-sentence passages as exemplified in (45). The items matched those used in Chapter 9, with the exception that the intervening *Also* sentence was removed.

(45) **Intro:** Bob suggested a business merger with Jennifer's company, and now she's considering it.

(i) **OT1H present:** <u>On the one hand</u>, she'd like to join forces with Bob, because he already has many loyal customers.

(ii) **OT1H absent:** She'd like to join forces with Bob, because he already has many loyal customers.

**OTOH:** <u>On the other hand</u>, she wants to make sure she can rise to power as CEO without competition.

The items were already normed for Chapter 9, so no norming pretest was conducted for the current study. For the current eye-tracking study, the 16 target items were interspersed with 96 items for an unrelated study. The proportion of experimental items in the study was 14%. Compare this proportion to that of Chapter 9: that eyetracking experiment included 36 items containing OTOH (24 experimental and 12 filler items) out of 84 items in total, meaning the proportion of OTOH items was 43%.

### 10.2.2 Dutch materials

The experimental stimuli consisted of 12 three-sentence passages, which were approximate translations of a subset of the items included in the eye-tracking experiment reported in Chapter 9. The experiment included 60 additional items for an unrelated study, which is why fewer experimental items were included in the Dutch experiment than in the English experiment. The proportion of experimental items in the Dutch eye-tracking experiment was 16.66%.

Two changes were made to the Dutch experimental items relative to those used in Chapter 9. First, the OT1H clause did not contain a subordinate *because* clause and was not followed by another intervening sentence, since the other items in the current experiment consisted of three simple sentences as well. Second, the names in the items were changed to noun phrases. Passage (46) shows an example of an experimental item.

(46) **Intro:** De ondernemer overwoog om zijn zaak te fuseren met een andere zaak.
*The entrepreneur was considering merging his business with another business.*

(i) **OT1H present:** Aan de ene kant kon hij profiteren van de klantenbasis van de andere ondernemer.
*On the one hand, he could profit from the client base of the other entrepreneur.*

(ii) **OT1H absent:** Hij kon profiteren van de klantenbasis van de andere ondernemer.
*He could profit from the client base of the other entrepreneur.*

**OTOH:** Aan de andere kant wilde hij graag als enige de leiding hebben in zijn zaak.
*On the other hand, he wanted to be the only person in charge of his business.*

The items were normed in a coherence judgment experiment to ensure that any effects found would not be due to differences in coherence. 16 participants rated the naturalness of stories on a scale of 1 to 5, with 1 being very unnatural and 5 being very natural. The 12 items were counterbalanced across two lists. Every participant rated the naturalness of one version of 12 stories, as well as 12 filler items. The fillers consisted of short stories that were either very unnatural or very natural, to create a spectrum of natural stories for assessing participants’ understanding of the rating scale. Unnatural filler stories contained discourse violations such as an incorrect connective given the context (e.g., ‘so’ where it should be ‘because’) or an unexpected turn of events (e.g., receiving a bonus after having lost an important judicial case). The results showed no difference between the coherence of stories with OT1H (average rating of 3.85) and stories without OT1H (average rating of 3.76) ($\beta$ = 0.09, SE = 0.25, t = 0.38, $p$ = .69).

### 10.2.3 Predictions

The studies reported in the current chapter test two competing hypotheses that provide explanations for the lack of a facilitative effect of OT1H in Chapter 9. First, it is possible that OTOH was easy to process in the earlier eye-tracking study (and therefore harder to facilitate) due to its frequency in the experiment. A second hypothesis is that OT1H does not have a strong facilitative effect in English because OTOH appears often without the prior foreshadowing of OT1H. Given that OT1H occurs more frequently with OTOH in Dutch than in English, it is hypothesized that the facilitative effect of OT1H is greater in Dutch than in English.

If the first hypothesis holds, the results from the English eye-tracking experiment will show shorter reading times on OTOH for stories also containing OT1H, compared to those that do not contain OT1H. This result would suggest that the frequency of OTOH in the earlier study concealed a possible facilitative effect of OT1H.

Evidence for the second hypothesis cannot be found by directly comparing the reading times between experiments, because they are conducted in different languages. However, if a greater effect of presence of OT1H is found in Dutch than in English, this can be considered evidence for our hypothesis. In order words, the current experiments will provide evidence for the second hypothesis if the Dutch eye-tracking study shows a more extensive facilitative effect across more eye-tracking measures compared to the English eye-tracking study. This result would suggest that comprehenders are indeed sensitive to language-specific features such as the distributions of markers in natural language.

Note that the two hypotheses are not mutually exclusive: it is possible that OT1H has a smaller facilitative effect in English than in Dutch, and that this facilitative effect was reduced even more by the large number of items in the previous experiment.

In order to investigate whether the Dutch and English items differ from each other in the predictions that contrast1 evokes (in the presence and absence of OT1H), a story continuation study was conducted. The results of this study will reveal whether the items are comparable in the absence of OT1H; that is, whether an unmarked contrast1 leads to the same predictions in the English as in the Dutch stimuli. This is crucial for the interpretation of the results in the eye-tracking studies: if the continuation study shows that items in the OT1H absent condition receive more contrastive continuations in one language over the other, it will indicate that any possible difference between the English and Dutch results could be due to the construction of the items, rather than distributional differences between the languages.

## 10.3 Experiment 1: Story continuation study

The goal of this study was to test whether the English and Dutch items differ from each other in the predictions that contrast1 evokes (in the presence and absence of OT1H). We asked native English and Dutch speaking participants to write a story continuation (one or two sentences) to the English or Dutch items, respectively, without the OTOH sentence.

### 10.3.1 Participants

40 native English speakers (age range 20–64 years; mean age 34 years; 31 female), and 40 native Dutch speakers (age range 18-61 years; mean age 29 years; 16 female), all registered as participants on the Prolific crowdsourcing website, received monetary compensation for their participation (1.2 GBP). Participants had various educational levels ranging from high school to a doctorate degree.

### 10.3.2 Materials and procedure

The experimental stimuli consisted of the 16 English and 12 Dutch items that were selected for the eye-tracking experiments. The items did not contain the OTOH sentence. The experiment was hosted on Lingoturk (Pusse et al., 2016) and distributed via Prolific.[1] Each version of every item was completed by ten people. Each participant saw one version of 8 English or 6 Dutch stories, and 10 filler items. They therefore only saw the marker OT1H four (English) or three (Dutch) times throughout the experiment. Filler items consisted of short stories in the same format as the experimental items, without the markers OT1H or OTOH.

### 10.3.3 Analysis procedure

The continuations were manually annotated for the presence of an explicit marker (i.e., whether a connective was used, and if so, which one), and discourse relation (i.e., whether the content of the continuation provided a contrast with the OT1H clause).

[1] www.prolific.ac

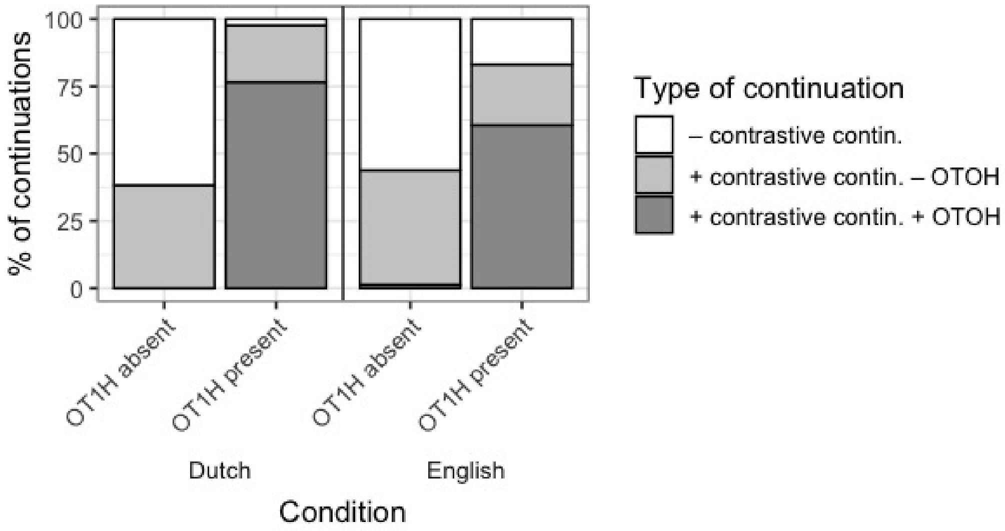


**Figure 10.1:** Percentage of contrastive continuations per condition and language.

### 10.3.4 Results

The results were modeled using binomial mixed-effects regression models. Contrastive continuations were modeled as a binary response variable; condition and language were deviation-coded and included as predictors. We started with a maximal random effects structure (random slopes for subjects and condition, and items and condition), but in order for the model to converge, the models did not include a random slope for item, nor a correlation between participant and condition. Significance of fixed effects was evaluated by performing likelihood ratio tests, in which the fit of a model containing the fixed effect for each condition is compared to another model without it.

Figure 10.1 presents the percentage of contrastive continuations (with and without the use of OTOH) per condition and language. As expected, both English and Dutch items received more contrastive continuations for conditions where OT1H is present than for conditions where OT1H is absent ($\beta$ = 3.72, SE = .42, $z$ = 8.75, $p$ < .001). More interestingly for the current study, the model shows an interaction effect of language and condition ($\beta$ = -2.59, SE = .77, z = -3.34, $p$ < .001): participants provided significantly more contrastive continuations for items where OT1H was present in Dutch than in English, but no such difference was found for items where OT1H was absent.

Finally, regarding the use of OTOH in Dutch and English, we note that OT1H was followed by OTOH specifically (rather than another connective such as *but*) in 77% of continuations in Dutch, and 62% of continuations in English. This difference between languages is not significant ($\beta$ = -0.89, SE = 0.58, z = -1.54, $p$ = .12); and the interaction between language and condition is only marginally significant ($\beta$ = -13.45, SE = 88.22, z = -0.15, $p$ = .07).

### 10.3.5 Discussion completion study

As expected, more contrastive continuations were provided for items where OT1H was present, compared to when it was absent. However, this effect is modulated by language. In Dutch, when OT1H is present, readers anticipate a contrast2 more strongly than in English, as indicated by a significantly higher proportion of contrastive continuations in Dutch compared to English. For items in which OT1H is absent, there is no difference between Dutch and English.

The results found in the current study do not confirm or deny the second hypothesis, which is based on the notion that OTOH occurs more frequently without prior foreshadowing of OT1H in English than in Dutch. However, these results do indicate that there is in fact a difference between the English markers and their Dutch equivalents. The Dutch pair seems to be more restricted: compared to English, Dutch "Aan de andere kant" (OTOH) is less likely to appear without having been foreshadowed with a prior "Aan de ene kant" (OT1H). In English, "On the other hand" does not necessarily follow "On the one hand".

The continuation of OT1H items with the marker OTOH also provides insight regarding the possible processing difficulty of OTOH in the absence of OT1H. If the distribution of the OT1H–OTOH pairing (i.e., how often OT1H is followed by OTOH) differs between English and Dutch, this might affect expectancy of OTOH in the context of OT1H. The results indicate that OT1H is followed by OTOH more often in Dutch than in English, but this difference is not significant. Any possible difference in processing time of OTOH in the upcoming English and Dutch eyetracking experiments can therefore not be attributed to the expectancy of OTOH in the context of OT1H (it would rather be due to the expectancy of OTOH in the absence of OT1H).

In sum, the results from the current study suggest that OT1H in the Dutch passages evokes a stronger expectation of OTOH than in the English passages. In the next two sections, we look at whether the presence of OT1H facilitates the processing of OTOH, and whether this effect is modulated by language.

## 10.4 Experiment 2: English eye-tracking study

An eye-tracking study was conducted to investigate the effect of the presence of the marker OT1H on the anticipation of OTOH. A relatively large number of participants was presented with a relatively small number of OTOH items, in order to prevent readers from getting accustomed to seeing many items containing OTOH.

### 10.4.1 Participants

80 native speakers of English (age range 18-41 years; mean age 20 years; 63 female) participated in this experiment. Data of 4 participants could not be used due to

problems with the eye-tracker. These data were removed before analysis. Participants were recruited from the Lancaster University community. All participants had normal or corrected-to-normal vision. Participants received course credit or monetary compensation for their participation. They were unaware of the purpose of the experiment.

### 10.4.2 Materials

The experimental stimuli consisted of the 16 three-sentence passages that were also used in the story continuation experiment. These items were interspersed with 96 items for an unrelated study. The stimuli were counterbalanced across two lists, with each story appearing in a different condition in each list. The participants were randomly assigned to one of the lists.

### 10.4.3 Procedure

Participants were tested individually. They were seated at a distance of approximately 60 cm from the monitor and rested their head on a chin-rest. Eye movements were recorded with SR Research Eyelink 1000 at a sampling rate of 500 Hz. The experiment lasted approximately 50 minutes.

Each session started with an oral instruction, after which the eye-tracker was adjusted if necessary. A brief calibration procedure was then performed, during which the participants had to fixate a random sequence of 9 dots at various locations on the screen. Upon successful calibration, the experiment started with two practice trials. The participant was instructed to read the passage at a natural pace and press the space bar after reading the entire story. Before presentation, a fixation mark appeared at the position of the first word of the first sentence. Participants were instructed to fixate this mark, after which the story appeared. The stories were presented randomly and in their entirety on the screen. The critical region (“On the other hand,”) never appeared at the beginning or end of a line. A verification statement about the story followed 25% of the items to ensure that the participants read the passages carefully. Participants indicated whether the statement was correct or incorrect by pressing a button on a button box.

### 10.4.4 Analysis procedure

For analysis purposes the sentences were divided into three regions, as illustrated in 47:

(47) (...) / famous customers. $_{pre-criticalregion}$ / On the other hand, $_{criticalregion}$ / she wants $_{spillover}$ / (...)

The critical region is where a possible facilitative effect of OT1H is expected to be found: if readers anticipate OTOH based on OT1H, reading time of OTOH should be faster in the OT1H present condition than in the OTOH absent condition.

Three reading time measures were computed: first pass duration, regression path duration and total reading time. First pass duration is the time spent in a region before moving on or looking back. This measure reflects the immediate processing difficulties a reader has when reading a region for the first time (Rayner, 1998). Regression path duration is the summed fixation duration from when the current region is first fixated until the eyes enter the next region on the right. This measure thus includes regressions to regions to the left of the current region. Regression path duration can be seen as reflecting the process of integrating the linguistic material with the previous context (Rayner, 1998). Total reading time is the total time spent in a region, including regressions to that region.

Prior to all analyses, skipped regions were treated as missing data. In all reading time measures, fixations shorter than 80 ms were removed. Outliers were removed by excluding reading times more than two standard deviations from both the participant's mean and the condition's mean in a region (1.5% of the data points for the first pass duration, 1.5% for the regression path duration, and 0.8% for the total reading time duration).

### 10.4.5 Results

Table 10.1 reports the mean reading time measures per condition and region, and Figure 10.2 shows the first pass duration per region. Reading times were modeled using generalized linear mixed-effects regression models with a gamma distribution. Subjects and items were included as random effects. In order for the models to converge, the random effects structure did not include a random slope for item and condition. Likelihood ratio tests were computed to compare models differing only in the presence or absence of the fixed factor for condition, which was deviation coded. The results of the models are summarized in Table 10.2.

At the critical region, the results showed a significant difference between the two conditions for first pass duration: the reading times on OTOH were shorter when OT1H was present compared to when it was not.

No significant effects were found at the critical region for regression path duration and total fixation duration. Moreover, no effects were found in the pre-critical region or spillover region.

### 10.4.6 Discussion English study

As predicted by the first hypothesis, whereby the frequency of OTOH in the earlier study concealed a possible facilitative effect, this study with the reduced number of OTOH items does show the predicted facilitative effect. The results indicate that

| | **Region** | | | | | |
|---|---|---|---|---|---|---|
| | *Pre-crit.* | | *Critical* | | *Spillover* | |
| | M | SD | M | SD | M | SD |
| *First pass durations* | | | | | | |
| OT1H present | 319 | 164 | 387 | 184 | 278 | 184 |
| OT1H absent | 317 | 169 | 399 | 180 | 259 | 134 |
| *Regression path durations* | | | | | | |
| OT1H present | 437 | 290 | 423 | 248 | 292 | 216 |
| OT1H absent | 453 | 297 | 442 | 244 | 271 | 158 |
| *Total reading time durations* | | | | | | |
| OT1H present | 385 | 229 | 436 | 235 | 300 | 203 |
| OT1H absent | 398 | 232 | 446 | 208 | 287 | 166 |

**Table 10.1:** Mean reading times and standard deviations in milliseconds per measure and region.

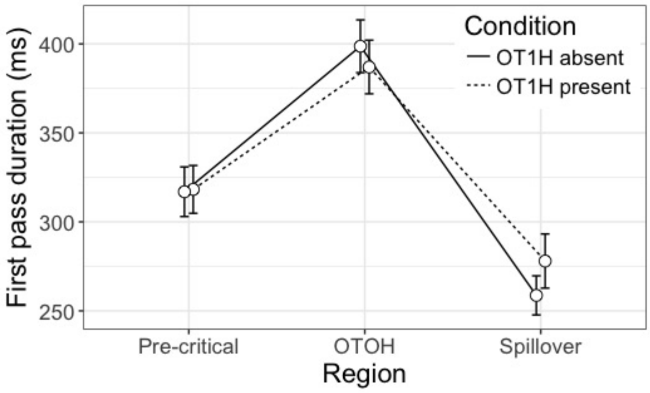


**Figure 10.2:** First pass duration per condition and region. The error bars represent the 95% confidence interval.

| | **Region** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Pre-critical* | | | | *Critical* | | | | *Spillover* | | | |
| | $\beta$ | SE | t | $p$ | $\beta$ | SE | t | $p$ | $\beta$ | SE | t | $p$ |
| First pass | -0.28 | 9.21 | -0.03 | .98[2] | -18.55 | 6.44 | -2.88 | **<.05**[1] | 4.33 | 5.44 | 0.80 | .43[1] |
| Regression path | -11.76 | 9.70 | -1.21 | .23[1] | -12.54 | 9.96 | -1.26 | .21[2] | 4.76 | 6.15 | 0.78 | .44[1] |
| Total fixation | -13.01 | 8.57 | -1.52 | .13[1] | -11.31 | 7.77 | -1.46 | .15[1] | 1.96 | 6.45 | 0.30 | .76[1] |

[1] Model did not contain random slope for subject.

[2] Random slope for subject did not contain correlation with condition.

**Table 10.2:** Regression coefficients and test statistics from linear mixed-effects models for the effect of condition for each measure and region.

the presence of OT1H facilitates the immediate processing of OTOH: the first pass durations of OTOH indicate that readers have more difficulty when they first see OTOH without OT1H compared to when they see it with OT1H. The absence of a significant effect of condition on regression path duration and total fixation duration indicates that readers do not have more difficulty integrating OTOH with the previous linguistic material when OT1H is absent compared to when it was present.

The results suggest that the absence of an effect of OT1H in the previous chapter was in fact because the experiment contained a large number of items with OTOH (32 items out of 92 items total), increasing the expectedness of OTOH overall and thereby reducing the expected facilitative effect of OT1H. However, it should be noted that the items used in this study differed from those in the earlier study: the intervening *Also* sentence was removed. We will return to this issue in the general discussion (Section 10.6). First, we look at whether the facilitative effect is modulated by language-specific factors, by conducting a similar experiment in Dutch.

## 10.5 Experiment 3: Dutch eye-tracking study

We conducted an eye-tracking study to investigate whether the facilitative effect of OT1H on the processing of OTOH can be replicated in another language, and whether this effect is modulated by language-specific features; specifically, whether the facilitative effect is stronger in Dutch, in which OTOH appears more rarely without the prior foreshadowing of OT1H.

### 10.5.1 Participants

84 native speakers of Dutch (age range 18-55 years; mean age 24 years; 71 female) took part in this experiment. Participants were recruited from the Utrecht University community. All participants had normal or corrected-to-normal vision. Participants were paid for their participation and were unaware of the purpose of the experiment.

### 10.5.2 Materials

The experimental stimuli consisted of the 12 items used in the story continuation experiment. These items were interspersed with 32 items for an unrelated study, and an additional 28 filler items.

The stimuli were counterbalanced across two lists, with each story appearing in a different condition in each list. All participants saw one version of all stories. The participants were randomly assigned to one of the lists.

### 10.5.3 Procedure

Eye-movements were recorded using an SR Research Eyelink 1000 at a sampling rate of 500 Hz. Participants were seated approximately 60 cm from the monitor. The experiment lasted approximately 30 minutes.

Similar to the English study, each session started with an oral instruction, after which the eye-tracker was adjusted if necessary. After successful calibration, the experiment started with two practice trials. The participant was instructed to read the passage in a natural pace and press the space bar after reading the entire story. The critical region never appeared at the beginning or ending of a line. The stories were presented randomly and in their entirety on the screen. A verification statement about the story followed 25% of the items to ensure that the participants read the passages carefully.

### 10.5.4 Analysis procedure

For analysis purposes the sentences were divided into the same three regions as in the English study; illustrated in 48:

(48) (...) / andere ondernemer. $_{pre-criticalregion}$ / Aan de andere kant $_{criticalregion}$ / wilde hij $_{spillover}$ / (...)

Prior to all analyses, skipped regions were treated as missing data. Fixations shorter than 80 ms were removed. In all reading time measures, outliers were removed by excluding reading times more than two standard deviations from both the participant’s mean and the condition’s mean in a region (2.9% of the data points for the first pass duration, 1.3% for the regression path duration, and 1.2% for the total reading time duration).

### 10.5.5 Results

Table 10.3 reports the mean reading time measures per condition and region, and Figure 10.3 shows the first pass duration per region. Reading times were modeled using similar linear mixed-effects regression models as used for the analysis of the English data. Table 10.4 displays a summary of the results of the model.

At the critical region, the results showed a significant difference between the two conditions in the first pass duration, regression path duration and total reading time duration: the reading times on OTOH were shorter when OT1H was present compared to when it was not. At the pre-critical region, the results showed a significant difference in the regression path duration: the reading times on the two words preceding OTOH were shorter when OT1H was present than when it was absent. No other significant effects were found in the pre-critical region or spillover region.

| | **Region** | | | | | |
|---|---|---|---|---|---|---|
| | *Pre-crit.* | | *Critical* | | *Spillover* | |
| | M | SD | M | SD | M | SD |
| *First pass durations* | | | | | | |
| OT1H present | 383 | 166 | 339 | 140 | 221 | 90 |
| OT1H absent | 379 | 162 | 390 | 158 | 223 | 102 |
| *Regression path durations* | | | | | | |
| OT1H present | 418 | 216 | 378 | 205 | 238 | 135 |
| OT1H absent | 459 | 270 | 429 | 217 | 248 | 157 |
| *Total reading time durations* | | | | | | |
| OT1H present | 454 | 243 | 420 | 231 | 269 | 167 |
| OT1H absent | 444 | 225 | 474 | 228 | 262 | 142 |

**Table 10.3:** Mean reading times and standard deviations in milliseconds per measure and region.

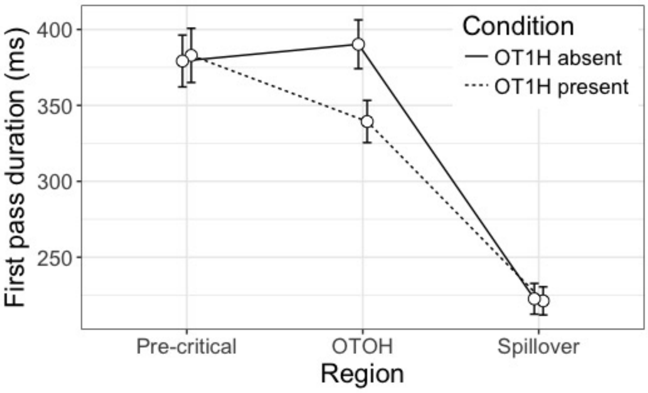


**Figure 10.3:** First pass duration per condition and region. The error bars represent the 95% confidence interval.

| | **Region** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Pre-critical* | | | | *Critical* | | | | *Spillover* | | | |
| | $\beta$ | SE | t | *p* | $\beta$ | SE | t | *p* | $\beta$ | SE | t | *p* |
| First pass | -4.13 | 11.10 | -0.37 | .71[2] | -49.15 | 10.54 | -4.67 | **<.001**[1] | -0.37 | 4.86 | -0.08 | .94[1] |
| Regression path | -34.21 | 10.84 | -3.16 | **<.01**[1] | -37.26 | 9.56 | 3.90 | **<.001**[2] | -4.51 | 9.37 | -0.48 | .63[1] |
| Total fixation | 2.16 | 11.71 | .19 | .85[1] | -61.43 | 10.16 | -6.05 | **<.001**[1] | -3.72 | 7.45 | -0.50 | .72[1] |

[1] Model did not contain random slope for subject.

[2] Random slope for subject did not contain correlation with condition.

**Table 10.4:** Regression coefficients and test statistics from linear mixed-effects models for the effect of condition for each measure.

### 10.5.6 Discussion Dutch study

As predicted by the second hypothesis, whereby the frequency of OTOH in natural language without prior foreshadowing of OT1H decreased a possible facilitative effect, this study on Dutch shows a more extensive facilitative effect across more eye-tracking measures: first pass duration, regression path duration, and total fixation duration on Dutch OTOH, as well as the regression path duration on the two words preceding OTOH, were shorter when OT1H was present compared to when it was not. These results indicate that Dutch OT1H allows readers to anticipate an upcoming contrast, which then facilitates the processing of OTOH. Moreover, the more extensive facilitative effect found in the current study compared to the English eye-tracking results suggest that cross-linguistic differences between the pair of markers do indeed influence the strength of the effect of OT1H on the processing of OTOH.

## 10.6 General discussion and conclusion

The studies reported in the previous chapter provided conflicting results: comprehenders were able to make fine-grained structural predictions of upcoming content based on OT1H, but no evidence was found that the presence of OT1H facilitated processing of OTOH. The studies reported in the current chapter aimed to further investigate this second finding. Two possible explanations were investigated. First, it is possible that OTOH was easy to process (and therefore harder to facilitate) due to its frequency in the experiment. We tested this hypothesis by conducting a follow-up eye-tracking experiment, in which a larger number of participants read a smaller number of items containing OTOH. A second hypothesis is that OT1H does not have a strong facilitative effect in English because OTOH appears often without the prior foreshadowing of OT1H. Given that OT1H occurs more frequently with OTOH in Dutch than in English, it was investigated whether the facilitative effect of OT1H is greater in Dutch than in English.

Regarding the first hypothesis, a story completion study showed that the marker OT1H does lead to predictions of upcoming contrast: participants provided more contrastive continuations to stories containing OT1H than to stories in which OT1H was absent. An eye-tracking study then revealed a facilitative effect of OT1H on OTOH: first pass durations on OTOH were faster for items in which OT1H was present compared to when it was absent. These results suggest that the frequency of OTOH in the earlier study did indeed conceal a possible facilitative effect.

Regarding the second hypothesis, the story completion study showed a difference in the number of contrastive continuations between languages: in Dutch, when OT1H is present, readers anticipate a contrast2 more strongly than in English. These results suggest that the Dutch pair is more restricted than the English pair. The Dutch eye-tracking study revealed a stronger facilitative effect of OT1H in Dutch compared to English: first pass durations, regression path durations and total fixation durations on OTOH were faster for items in which OT1H was present compared to when it was absent. Additionally, regression path duration of the two words preceding OTOH was faster if OT1H was present in Dutch, but not in English. These results suggest that the facilitative effect of OT1H is indeed modulated by the distributional characteristics of the pair of markers: OTOH occurs more frequently with OT1H in Dutch than in English.

The results from the current studies therefore suggest that both hypotheses hold. However, before one can fully accept the hypotheses, an alternative explanation for these results should be considered. Both the English and the Dutch materials differed from the materials used in the earlier studies because they were interspersed with other items for unrelated studies. These adaptations could have had an effect on the results found in the current experiment. For the English experiment, the intervening *Also* sentence was removed from the experimental items. It is possible that this difference can (partly) explain for the lack of a facilitative effect in the earlier study: the intervening material could have weakened the effect of OT1H on processing of OTOH. A similar observation is made for the comparison between the English and Dutch studies in the current chapter: the OT1H-sentences in the Dutch study did not contain a *because* clause, whereas those in the English study did. Again, it is possible that the intervening clause had an effect on readers’ expectations.

However, it should be noted that a growing body of research suggests that readers can maintain expectations across adjacent sentences (e.g., Drenhaus et al., 2014; Kehler et al., 2008; Rohde & Horton, 2014; Xiang & Kuperberg, 2015), and the studies reported in the previous chapter showed that they can even be maintained across multiple sentences. We therefore argue that the adaptations cannot solely account for the effects found in the current eye-tracking experiments.

Nevertheless, the question as to what the “half-life time”[2] of expectations is provides an interesting avenue for future research. This can be addressed by investigating whether readers are able to maintain expectations across larger spans of discourse. The OT1H/OTOH pair provides an ideal testing ground for this, but it would also be interesting to test the generalizability of the discourse prediction effects observed here using other text structuring markers such as list signals (e.g., *There are three things that are relevant. First... Second... Third...*). These might be particularly interesting because their distribution in a text (holding between larger text segments) might differ from the distribution of OT1H / OTOH.

In sum, the results from the current studies suggest that OT1H in English does facilitate the processing of OTOH, but not as strongly as in Dutch, where the pair of markers occurs together more frequently. Comprehenders therefore seem to be sensitive to distributional differences in languages, which can affect the amount of surprisal that a comprehender experiences on a discourse marker: OTOH occurs without OT1H more frequently in English than in Dutch, which makes English OTOH

[2] Proposed by Torgrim Solstad, pc.

easier to process when OT1H is absent. The findings from these studies hence suggest that the effect of discourse markers on comprehenders' processing is modulated by language-specific factors, and can therefore differ cross-linguistically. These results highlight the importance of cross-linguistic research, and will hopefully provide input for future studies investigating the effect of discourse markers on processing. One interesting question is whether the facilitative effect of the first marker in a pair of discourse markers can be replicated in Chinese, in which different pairs of markers (such as *because/then* and *although/but*) occur together frequently (Steele & Specia, 2014; Xue, 2005). Research into phenomena such as these can provide more insight into the effect of language-specific factors on the processing of discourse structure, as well as a better understanding of discourse processing in general.

# Chapter 11

# Discussion and conclusion

Understanding a discourse requires comprehenders to infer coherence relations between clauses and sentences in a text. If they are not able to construct such relations, they will fail to fully understand that text. Discourse coherence is therefore crucial to natural language comprehension, and a more detailed understanding of how discourse coherence is represented and processed would contribute to research on natural language in general.

This thesis aimed to achieve a better understanding of the distinctions between coherence relations by focusing on three challenges: (i) the differences between various theories, and between the types of coherence relations they propose, (ii) patterns of discrepancies between annotations, and (iii) interpretations of ambiguous relations. Issues related to these challenges were investigated using a combination of theoretical exploration, corpus-based methods, and experimental methods. The current chapter summarizes the main findings of the thesis, and points out directions for further research.

## 11.1 Overview of main findings

Much of the research in the field of discourse coherence has focused on detailing a fixed set of coherence relation types, and on investigating the acquisition, production, processing and representation of such relations by language users. These efforts have provided valuable insight into the nature of coherence relations. For example, researchers have agreed on the existence of coarse-grained relation types such as causal and adversative relations, and studies have shown that these types are cognitively relevant (i.e. they are acquired, represented and processed differently). However, there are still many questions left unanswered regarding the nature of coherence relations, such as: which finer-grained distinctions of coherence relation types are cognitively relevant? Between which types of segments do comprehenders infer relations (e.g., adjacent sentences, non-adjacent sentences, paragraphs, etc.)? Do comprehenders construct expectations about upcoming relations? How are coherence relations marked and what influences their marking?

This thesis addressed some of these questions by providing more insight into differences and similarities between relation types distinguished in frameworks, which types of relations comprehenders construct and how these relations are inferred and processed. The current section gives an overview of the findings in relation to the research goals introduced in Chapter 1.

As discussed in Chapters 2 and 3, various frameworks have proposed inventories of coherence relations that differ in size and labels. The differences between the proposed inventories make it difficult to draw comparisons across the corpora annotated in the various frameworks. The first goal of the thesis was to explore the differences and similarities between these proposals in order to increase the interoperability between them. The goal was addressed in Chapter 3 by proposing the Unifying Dimensions (UniDim) approach, which can be used to translate relation labels from one framework into those of another framework. It therefore functions as an interlingua between various approaches.

The UniDim approach proposes mappings between relational labels. These mappings were created by decomposing all labels from the PDTB and RST-DT into their values on several dimensions and features that relations share. The starting point for UniDim was the Cognitive approach to Coherence Relations (Sanders et al., 1992).

Chapter 3.4 discussed the benefit of using a decomposition approach: by decomposing the relational labels, every label can be mapped to the closest corresponding match in another framework, even when no exact corresponding label exists. In cases where one framework distinguishes a type of relation that another framework doesn't, the interlingua indicates which labels in the "underspecified" framework are the most likely counterparts, given their similar specification on other features. By being able to identify equivalent labels in other frameworks, researchers working in different paradigmata could make use of all the annotated data that is available. Additionally, we hope that the interlingua will facilitate discussion between researchers working with different frameworks.

Given that the decomposition of relation labels was based on the definitions that the frameworks provided in their manuals, the Unifying Dimensions approach needed to be validated using annotated data. This was the second goal of the thesis. In order to validate the theoretical mapping, we compared annotations of different frameworks on the same data, to ensure that the provided annotations correspond to each other. Two combinations of frameworks were investigated, as reported in Chapter 4: (i) we annotated a corpus of spoken data using PDTB 3.0 and CCR, and (ii) we compared the annotations of newspaper articles of the Wall Street Journal that are contained in both the PDTB 2.0 and RST-DT. These studies did not only contribute to the evaluation of the mapping, they also represent the first exploration of the compatibility between relation label annotations from different frameworks. Moreover, the spoken data corpus can contribute to domain adaptation for automatic coherence relation classifiers by providing more training data from a different genre.

The results of the two efforts showed that the Unifying Dimensions approach resulted in a relatively accurate mapping between the various frameworks, but both studies also revealed several patterns of discrepancies between the annotations. Some of the disagreements were caused by different operationalizations of the frameworks, which in turn can be attributed to theoretical differences. For example, as discussed in Chapter 4.2, instances that were annotated as List relations in RST-DT were often annotated as Conjunction in PDTB, despite the PDTB having a corresponding List label available. Closer consideration of the manuals revealed that PDTB employs a stricter definition of List relations: in PDTB, lists have to be announced in the context, but in RST-DT, this is not a necessary requirement.

A second cause for the discrepancies between the frameworks was due to certain types of relations being able to express multiple readings. For example, many instances annotated as Instantiation and Specification in PDTB were annotated as Evidence or Explanation-argumentative in RST-DT. This finding raised several questions: what causes such discrepancies? Is this related to the multi-interpretability of relations? How can these discrepancies be evaluated? Are both annotations justified? How can this be investigated?

The third goal of the thesis was to address the issue of justification of the distinctions made in coherence relation approaches. Approaches are often justified by focusing on either the descriptive adequacy or the cognitively plausibility of distinctions. Descriptive adequacy relates to the intuitions and interpretations of researchers: descriptively adequate distinctions are tools that researchers can use to describe texts. Cognitively plausible distinctions, by contrast, are relational categories that play a role in the construction of a mental representation of a text.

One issue with descriptive adequacy is that descriptively adequate distinctions cannot be falsified. As discussed in Chapter 5.1.3, the descriptive adequacy of distinctions can be justified by successfully applying the theory to describe and analyze the structure of a text. However, the analysis of a text can be subjective, and often multiple ways of analyzing a text can be proposed. Descriptive adequacy does not provide any other external measures to verify or falsify the distinctions made in a theory. Falsification can be done, however, using empirical results from cognitive research, which speaks to the cognitive plausibility of the distinctions. But the literature does not provide many details regarding what types of empirical results or cognitive research would suffice. This was addressed in Chapter 5.

Chapter 5 proposed a method that details how distinctions between relations can be tested, in order to validate or falsify them. This validation approach is based on the notion that coherence relations are cognitive entities that people infer when constructing a mental representation of discourse. A "checklist" of sources is provided, which can be used to systematically evaluate the probability of a distinction being cognitively plausible. This method will hopefully contribute to the development of general theories of coherence.

A crucial source of evidence for the cognitive plausibility of relational distinctions comes from studies investigating readers' interpretations of which type of relation holds between two segments of text. However, asking naïve readers to provide a relation label for instances is not possible due to the complexity and meta-linguistic nature of the task: it would require training to get readers acquainted with the meaning of the labels that belong to an inventory. We therefore needed a method that can elicit readers' interpretations of which coherence relation holds between two segments of text. This formed the fourth goal of the thesis: to develop a new methodology that can be used to crowdsource interpretations of coherence relations. Chapter 6 describes the development of a connective insertion method, whereby participants are presented with segments and asked to insert a connective from a predefined list. The items that were used in the experiment reported in Chapter 6 came from Wall Street Journal texts that carried both PDTB and RST-DT annotations. We were therefore able to compare the insertions from our participants with the annotations from the corpora. The results showed that the majority of the inserted connectives converged with the original label. Further, the distribution of connectives revealed that multiple senses can often be inferred for a single relation. This method is thus suited to reflect the (possible) multiple readings of coherence relations.

We also explored an optimal design of the task by investigating how much context is most beneficial. In one condition, participants were presented with items without their linguistic context; in the other, participants saw two sentences preceding and one sentence following the item. The results showed no significant difference in distributions of insertions between conditions overall. This further supports the reliability of the method: the no-context condition showed a near perfect replication of insertions in the context condition. A by-item comparison did reveal that for a small portion of the items, the presence of context did influence the annotations when the fragments contained at least one of the following characteristics: (i) the context introduced the topic, (ii) the context sentence following the relation expanded on the second argument of the relation; or (iii) the first argument of the relation was a subordinate clause that attached to the context. The presence of context led to less agreement when the connection between the context and the first argument was not strong due to a paragraph break or a topic change (see also Chapter 6.4). These results provide fundamental insights into readers' sensitivity to context and the (limited) influence of context on the disambiguation of relations. Moreover, the knowledge regarding the effect of context can be useful in the design of coherence relation annotation tasks.

The connective insertion task provided a methodology to further investigate the discrepancy between PDTB and RST-DT annotators regarding the annotation of PDTB Instantiation and Specification relations. For a large number of these instances, RST-DT annotators coded an Evidence or Explanation-argumentative relation. This discrepancy can be explained by the fact that examples and specifications can often have multiple functions (cf. the multi-level thesis, see also Chapter

7.2): to illustrate/specify a situation, and to serve as an argument for a claim. The fifth goal of the thesis was to investigate the discrepancy regarding INSTANTIATION and SPECIFICATION relations by considering the interpretations of naïve readers. This was addressed in Chapter 7. The results show that these relations are indeed interpreted as ideational (INSTANTIATION and SPECIFICATION) and argumentative (EVIDENCE and EXPLANATION-ARGUMENTATIVE). These findings suggest that examples and specifications can have multiple, simultaneous readings. Both PDTB and RST-DT annotations are therefore justified (although neither annotation expresses this double reading).

The next question that arose is whether readers show systematic biases when interpreting such multi-interpretable relations. This question was investigated in Chapter 8, using the connective insertion method in a repeated measure design: the total set of items was divided into four smaller batches, and subjects were asked to participate in each of these over the course of a few months. The results showed that individual readers have consistent biases in how they interpret INSTANTIATION and SPECIFICATION relations: some readers are more prone to interpret relations as argumentative, whereas others are more prone to interpret relations as non-argumentative.

We investigated whether this difference in bias was caused by differing levels of processing (cf. the deep vs. shallow processing account, see Chapter 8.4): the same participants took part in a follow-up study in which they were asked to summarize the segments of relations using their own words (to ensure deeper processing), and then insert a connective. The results showed that the effect of prior bias was modulated by the task: participants with an argumentative bias interpreted fewer relations as argumentative when processing the relations deeper. Participants with a non-argumentative bias did not show a difference in their distribution of insertions between tasks. This suggests that shallow processing is related to inferring an argumentative reading, whereas deep processing is related to inferring a non-argumentative reading. These studies provided novel insights into the effect of individual variability on discourse interpretations.

The final goal of the thesis relates to the interaction between connectives and the content of the segments. Connectives are important signals of coherence relations: research has shown that they function as processing instructions during reading, by signaling to comprehenders how to link the segments. In other words, it is assumed that readers process both the connective and the content of the segments to infer a relation. However, the mappings in Chapter 4 indicate that comprehenders (or at least annotators) rely heavily on connectives, rather than the content of the segments, to infer a relation: in absence of an explicit connective, annotators show little agreement on the type of relation they infer. The sixth goal of the thesis was therefore to investigate whether readers make use of both the connective and the content of the segments when processing discourse. Furthermore, the effect of distributional differences of connectives in different languages on comprehenders' processing was investigated, in order to explore whether discourse expectations are modulated by language-specific factors.

The cue phrase *On the one hand* (OT1H) provided an ideal testing ground for investigating whether readers take into account both the connective and the content of the segments during processing: OT1H raises an expectation of an upcoming contrast, which can be satisfied by a sentence marked by various markers. Chapter 9 presented studies for which items contained a sentence marked by OT1H, which presented contrast1, and a sentence marked by *On the other hand* (OTOH), which presented contrast2. Additionally, an intervening sentence marked by *But* occurred between these two sentences (see also Chapter 9.3). In the global condition, this sentence presented contrast2 (and therefore satisfied the prediction set up by contrast1); in the local condition, it did not present contrast2 (and therefore did not satisfy the prediction for upcoming contrast). Three experiments using offline and online measures show that comprehenders disprefer a subsequent contrast marked with *On the other hand* when the intervening content established an appropriate contrast with *On the one hand*. This indicates that comprehenders make use of both the connective and the content of the segments to infer relations during processing.

An additional condition was added to investigate whether the presence of OT1H facilitated the processing of OTOH. The results from the eye-tracking study did not reveal such a facilitative effect. Two possible explanations were further investigated in Chapter 10. First, we hypothesized that a possible facilitative effect of OT1H was reduced because the expectedness of OTOH overall increased throughout the study, due to a large number of experimental items. An English eye-tracking study containing fewer OTOH items revealed a facilitative effect of OT1H on OTOH. These results suggest that the frequency of OTOH in the earlier study did indeed conceal a possible facilitative effect. A second hypothesis was that OT1H does not have a strong facilitative effect in English because OTOH appears often in natural language without the prior foreshadowing of OT1H. Given that OT1H occurs more frequently with OTOH in Dutch than in English, it was investigated whether the facilitative effect of OT1H is greater in Dutch. A Dutch eye-tracking study revealed a stronger facilitative effect of OT1H in Dutch compared to English. The findings from these studies hence suggest that the effect of discourse markers on comprehenders' processing is modulated by language-specific factors such as distributional characteristics, and can therefore differ cross-linguistically.

## 11.2 Directions for future research

The results obtained in this thesis open various directions for further research. This section points out some interesting future directions which build on the present work. Some of these have already been brought up in earlier discussions, but are summarized here to provide a clear overview of possible research directions.

### 11.2.1 Future work on annotation projects

**Agreement on annotation of implicit and explicit relations** Annotated data is an important resource for the linguistics community, which is why researchers need to be sure that such data are reliable. However, arriving at sufficiently reliable annotations appears to be an issue within the field of discourse coherence, possibly due to the fact that coherence is a mental phenomenon rather than a textual one. Spooren & Degand (2010) and Hoek & Scholman (2017) provided advice on how to evaluate coherence relation annotations, and they advocated for a more transparent discussion of the annotation process, data, and outcome. The current thesis contributes to this discussion by providing insight into the agreement on implicit relations. Specifically, the PDTB–RST-DT mapping discussed in Chapter 4 indicated that agreement on implicit relations is lower than on explicit relations (see also Sanders et al., 2018). Intuitively this makes sense, since implicit relations contain no or less explicit linguistic markers on which coders can base their decision. Reports of annotation efforts currently provide agreement statistics for implicit and explicit relations combined, but this conceals the agreement on implicit specifically, which can be assumed to be lower. In order to be able to estimate the scope of this issue (i..e., lower agreement on implicit relations), we recommend future annotation efforts to be more explicit about the agreement on implicit and explicit relations separately.

Future work can explore this presumed difference in agreement on implicit and explicit relations. Specifically, what causes the disagreement on implicit relation annotation (besides the obvious lack of connectives)? Is it that annotators are not aware enough of other signals? Or is it that instances can express several readings, and the disagreement between annotators actually reflects the distribution of relational senses – that is, the disagreement is in fact reflective of a relation's multi-functionality? If the difference is caused by the increased ambiguity of implicit relations, one solution would be to require annotators to annotate multiple relational labels (if possible) for every relation.

It would also be interesting to investigate the agreement on relations marked by underspecified connectives such as *and* and *but* (see also Spooren, 1997; Asr & Demberg, 2012). Such underspecified connectives signal a relation that is less specific than the one that readers will actually construct. For example, the temporal connective *after* does not signal the causal reading that comprehenders will likely infer for Example (49).

(49) *Theo beamed* <u>after</u> **his boss Joanne complimented his work.**

Relations marked by underspecified connectives can therefore be considered more similar to implicit relations than to explicit relations (see also Hoek, 2018). It would be interesting to investigate whether agreement on the annotation of such underspecified relations is also lower than on explicit relations.

Finally, the marking of coherence relations by other linguistic elements deserves more investigation. Connectives have been at the center of research in the signaling literature, but corpus data has shown that many relations are actually not signalled by a connective (between 20-50%, see Das & Taboada, 2018; Prasad et al., 2007). The RST Signalling Corpus (Das & Taboada, 2018) presents the most elaborate research effort to identify other relational signals. They identified linguistic cues that signal coherence relations annotated in the RST-DT. However, they did not link signals to relational types in a systematic way, and it is not clear how or why the indicated signals function as cues for coherence relations. Hoek (2018) addressed this by identifying three distinct ways in which segment-internal elements can systematically interact with connectives to express a relation. They also provide examples of elements that do this, such as negation words and verb tense, and evaluate this using data from a translation corpus. The current thesis took a different approach to identifying relational signals. Specifically, Chapter 7 identified features that relations interpreted as ideational or argumentative had in common. Hence, this approach took into account readers' interpretations as well as corpus data. All three lines of research provide interesting insights into relational signals, and can be expanded on in future research. By gaining additional insight into typical markers of relations, the agreement on coherence relation annotations can hopefully be improved.

**Automatic identification of relations** Manual annotation is a costly and time-consuming procedure. In the future, this procedure can hopefully be replaced by automatic discourse relation classifiers. The classification of explicit relations into coarse-grained classes is relatively easy and accurate (see, for example, Pitler et al., 2008), but the agreement on fine-grained relational classes for relations marked by ambiguous connectives such as *but* is more problematic. Moreover, the classification of implicit relations needs improvement: two CoNLL shared tasks (Xue et al., 2015, 2016) report accuracies just over 40% F-score on implicit relation sense labeling for an 11-way classification.

The automatic classification of relations can be improved by having more training data (i.e., manually annotated data) available. The connective insertion method proposed in Chapter 6 provides a promising alternative to traditional annotation, and might therefore be used to generate large amounts of coherence relation annotations. Additionally, this method might contribute to the improvement of classifiers by providing less noisy training data: the connective insertion task provides a distribution of relational senses for every item, which reflects the true meaning of a relation better than a single label assigned by an annotator according to a specific framework. Given that coherence relations can be ambiguous and express multiple meanings, the additional meaning that such a distribution provides can improve the classification of both explicit and implicit relations. Future research can focus on how classifiers can be trained using a distribution of relational labels for every item.

However, more research is necessary regarding the optimal design of the connective insertion task. In its current form, relations from six coarse-grained relational classes can be distinguished. Future studies can focus on expanding the method so that more types of relations can be annotated. The method can then also be used to generate annotated data from other genres, which can contribute to domain adaptation.

We are currently experimenting with a two-step annotation approach, whereby annotators are free in their choice of connective in the first step (that is, no choices are given), and are asked to disambiguate their choice using the drag-and-drop interface with predefined connectives. The connectives will be generated based on the input in the previous step. This design prevents participants from being overloaded with connectives in a single step. Results from this experiment will provide more insight into the feasibility of the connective insertion method for large-scale discourse annotation efforts.

### 11.2.2 Future work on interpretation and processing experiments

**Individual differences** Prior to the work reported in this thesis, few studies had investigated how readers actually interpret multi-interpretable coherence relations: do they have consistent interpretation preferences, and if so, do they show individual variability in these preferences? Chapter 8 showed that comprehenders indeed have consistent biases for interpreting the ideational or argumentative function of INSTANTIATION and SPECIFICATION relations. This finding raises several questions regarding individual variability in coherence relation interpretation and will hopefully trigger follow-up research in the community. For example, future research can investigate whether such individual interpretation preferences also occur for other types of relations. Moreover, how do individual differences affect coherence relation interpretation and processing in general? What causes such differences? And how do these differences interact with theories of general interpretation strategies, such as the causal preference theory (Sanders, 2005, 2017) or the continuity hypothesis (Murray, 1997)?

The deep vs. shallow processing account also provides an interesting avenue for future research. Several studies indicate that comprehenders' reading processes are affected by the depth of processing (e.g., Aaronson & Ferres, 1986; Hannon & Daneman, 2004; Just & Carpenter, 1978; Lehman & Schraw, 2002; Noordman et al., 1992). The results suggest that readers are satisfied with a rather *parsimonious* (i.e. economical or shallow) processing in normal reading, but when they read the text with a particular goal in mind (e.g., retaining more information), they process sentences more thoroughly and engage in more inference processes. However, as raised in the previous paragraph, we do not know yet what constitutes parsimonious processing with respect to coherence relations: do readers have a "default" relation type that they infer when they process text on a shallow level?

Moreover, the results reported in Chapter 8 suggest that comprehenders differ in their default processing depth: the distribution of insertions for readers with a non-argumentative bias did not change between tasks, indicating that their processing level was deep in both tasks. Readers with an argumentative bias did show a different distribution (more non-argumentative insertions) when the task required them to process more deeply. It would be interesting to investigate whether this difference in processing depth was typical of the connective insertion task (e.g., some participants found the task easier to do than others), to crowdsourcing studies (e.g., some are more invested in providing quality work than others), or perhaps to reading in general. If the latter turns out to be true, what causes this general variability in reading?

**Representation and processing of various types of relations** As discussed in Chapter 5, there is ample evidence for the representation and processing of coarse-grained relation types (such as causal versus additive or adversative), but little research has been conducted regarding other types of relations. This has resulted in a gap in our knowledge of discourse coherence in relation to cognition: how do certain types of relations influence representation and processing? Future work will hopefully address this by investigating relation types other than causal relations. For example, it can be investigated whether SYNCHRONOUS and ASYNCHRONOUS temporal relations have different effects on processing (i.e., is one generally processed faster than the other?). Given that the PDTB and RST-DT disagree on the annotation of LIST relations (see Chapter 4), it can be investigated whether LIST relations are in fact processed differently from CONJUNCTION relations. Similarly, are comprehenders able to distinguish between the various types of CONDITION relation types that PDTB and RST-DT propose, or between the types of RST-DT ELABORATIONS? This could be investigated using a sorting task (see Chapter 5). Considering the relatively limited amount of research on the interpretation and processing of non-causal relations, many more research questions like these can be formulated and would provide more insight into the relational constructs that are present in our linguistic system.

**Segment size and expectations** Most of the research on discourse expectations conducted in the field so far focused on expectations regarding relations within sentences or between adjacent sentences. Chapter 9 contributed to this line of research by providing evidence that readers are also able to construct and maintain expectations of upcoming coherence relations across larger text spans. This addresses the issue of segment size: many coherence relation approaches differ in the size of segments that they consider to be relational arguments. Although no approach would exclude explicit relations with segments that span across multiple sentences, few studies have investigated whether and how comprehenders infer relations between non-adjacent sentences. An interesting avenue for future research is therefore to investigate between which text spans comprehenders are able to construct coherence relations. Are comprehenders able to construct non-local relations, or does this come at a processing disadvantage? How well can they deal with sentences intervening between the segments, and how much intervening material can they handle without experiencing processing difficulties?

In a similar line of reasoning and relating to the findings in Chapter 10, an interesting topic for future research would be to investigate what the "half-life time" of expectations are: when comprehenders construct expectations of upcoming coherence relations, how long do they maintain these expectations? If these expectations are not maintained for long periods of time, is this because the comprehender gives up on them, or because these expectations are "overridden" by expectations generated based on the incoming input? Does the half-life time of expectations depend on readers' individual characteristics, such as working memory?

These questions can be addressed by replicating the experiments reported in Chapters 9 and 10 with more non-contrastive intervening material between contrast1 and contrast2. Another approach would be to investigate this issue using different relational types, such as LIST relations (which can be expressed by markers such as *First... Second...*).

**Context and contextual signals** Prior research that has investigated how coherence relations are marked has mainly focused on the function of connectives in discourse. More recent efforts, however, have focused on identifying additional signals, but these signals tend to be segment-internal (see, for example, Hoek, 2018; Das & Taboada, 2018). Much less is known about the influence of context on the interpretation of coherence relations. Related to this, approaches also seem to differ in how they treat context. For example, Chapter 4 highlighted the difference in operationalization of LIST relations in PDTB and RST-DT: PDTB requires them to be announced in the context, whereas RST-DT does not. This difference raises the question of whether readers are in fact sensitive to other types of regularities that might elicit relational expectations (referred to as "relational cues" here) and are located outside of the relation, in the linguistic context.

The evidence for readers' sensitivity to contextual cues is mixed. Sanders (1997) showed that readers took into account the context for determining the subjectivity of coherence relations, but only for those relations that were ambiguous. Canestrelli et al. (2016) did not find evidence that contextual evaluative (subjective) markers facilitated the processing of subjective causal relations. Chapter 6 also did not reveal a significant difference in interpretations of relations presented with or without their linguistic context. However, this might be caused by the fact that the context and items were not manipulated: it could be that there were no strong relational cue words in the context of enough experimental items to reveal an effect.

The influence of context on the interpretation of coherence relations deserves more consideration. Future research can investigate what type of relational cues actually occur outside of the relation, and how often this occurs. Another interesting line of research is to explore how sensitive readers are to such cues that occur outside of the segments. Currently, we are investigating whether readers are sensitive to LIST cues (such as *a few* and *several*) that precede the segments of a relation, and whether the sensitivity is influenced by individual differences in working memory and linguistic experience.

**Multi-level thesis** The multi-level thesis stipulates that given two discourse segments, there are two levels at which a relation can hold (Moore & Pollack, 1992, see also Sanders & Spooren, 1999). Theoretical accounts often agree on the existence of both of these functions, but not on how the functions relate to each other, and whether these functions exist for *every* (type of) relation. Few studies have experimentally addressed readers' processing and representation of these levels of discourse. Chapters 7 and 8 represent some of the first efforts to do so. The results revealed that the systematic disagreement between the PDTB and RST-DT regarding INSTANTIATION and SPECIFICATION relations can be attributed to the fact that coherence relations can function on an ideational and argumentative level. One interesting line of future research would therefore be to investigate which other types of relations can function on different levels.

Moreover, an open research question is whether comprehenders can actually interpret these different functions simultaneously, and if so, whether this has an influence on the processing of such relations. It is possible that multi-functional relations require more processing effort because comprehenders can construct different interpretations. However, it is also possible that comprehenders in fact commit to one interpretation early on during processing, in which case the processing of multi-functional relations should not require more effort.

### 11.2.3 Future work on the development of cognitively plausible theories of coherence relations

**Cross-linguistic research** In order to advance our understanding of discourse coherence, we need cognitively plausible theories that have explanatory power. These theories can be used to generate hypotheses that can then be tested in order to try and falsify them. As discussed in Chapter 5, it is generally assumed that cognitively plausible theories are not language-specific, but rather should generalize over all languages. However, this assumption deserves more investigation, given that cross-linguistic studies have revealed significant differences as well as similarities between the ways in which coherence relations are realized, distributed and interpreted across languages. For example, as discussed in Chapter 5, subjectivity is realized on the connective level in many languages (e.g., Dutch, German, French and Chinese have connectives that prototypically mark either objective or subjective relations), but not in English. It is possible that subjectivity is realized on other levels in English, however, such as pitch. Another example comes from studies in Chapter 10, which indicated that comprehenders are sensitive to distributional characteristics in languages, and that the effect of discourse markers on comprehenders' processing is modulated by such language-specific factors.

A starting point to investigate cross-linguistic differences in relation types could be to explore the fine-grained adaptations to the PDTB relational inventory for languages such as Arabic (Al-Saif & Markert, 2010), Italian (Tonelli et al., 2010), and Chinese (Zhou & Xue, 2015) (see Chapter 5). Are these adaptations caused by differing expert intuitions (relating to the descriptive adequacy of the distinctions), or by actual linguistic differences (relating to the cognitive plausibility of the distinctions)? One crucial aspect to consider is the influence of non-linguistic experience: do specific cultures have a "need" for different types of relations to be expressed? Exploring these issues will provide more insight into the question of whether relational categories can in fact be language-specific.

**Grain size** Most proposals for relational inventories agree on the more coarse-grained classes of relations, such as CAUSE, TEMPORAL and ADVERSATIVE. As discussed in Chapter 5 and shown in Appendix E, there is ample evidence for the cognitive plausibility of these coarse-grained types. However, more research is needed regarding the cognitive plausibility of fine-grained relation types. How fine-grained are the relational constructs that comprehenders can infer? For example, are comprehenders able to distinguish between volitional and non-volitional causal relations when interpreting and processing discourse, or do they only distinguish between more coarse-grained types such as causal versus non-causal relations (see also Kaiser, 2012)? Moreover, do comprehenders differ in the grain size that they can construct?

The linguistic system provides us with specific connectives for expressing fine-grained relations, which supports these distinctions in annotation work. It remains an open question whether such fine-grained distinctions are also relevant in processing. For example, do various types of ELABORATIONS (such as PART-WHOLE and SET-MEMBER in RST-DT) have different effects on comprehension and processing? Are comprehenders still able to distinguish between such fine-grained relation types, or are they only able to distinguish between more coarse-grained relation types?

If future research shows that fine-grained relation types that do have unique corresponding connectives are not as relevant in comprehension and processing, what does this mean for the cognitive plausibility of the distinctions? Consider SPECIFICATION and GENERALIZATION relations: they are similar in conceptual meaning, but they both have specific corresponding markers (*more specifically* and *more generally*, respectively). If future studies reveal no difference in production, acquisition, representation, processing and annotation of these fine-grained relation types, can we consider their probability of cognitive plausibility to be low (cf. Chapter 5)? Some might argue that the existence of a corresponding cue phrase is enough to justify such fine-grained distinctions. These issues require more consideration, research and discussion within the community before they can be solved.

## 11.3 Conclusion

The work reported in this dissertation focused on distinctions between types of coherence relations by comparing approaches, annotations, and interpretations. The studies provided insight into the similarities and differences between the inventories of coherence relations that are proposed by the approaches, as well as the operationalizations of these approaches. We then revealed how these factors can affect the resulting annotations. By examining how different types of relations are interpreted and processed by naïve readers, this dissertation provided new insights into the processes involved in discourse comprehension.

This dissertation contributed to our understanding of discourse coherence, but it still leaves a large number of open questions for future research. The continued study of discourse coherence will bring us a step closer to natural language understanding by improving the modeling of the coherence relation constructs that are part of language users' mental representations. Such new insights can only be achieved by considering evidence from various sources, combining methodologies that are typical of text linguistics and psycholinguistics. Converging evidence can help us get a better understanding of the exact processes involved in discourse production and comprehension.

# Appendices

# Appendix A

# PDTB tagset

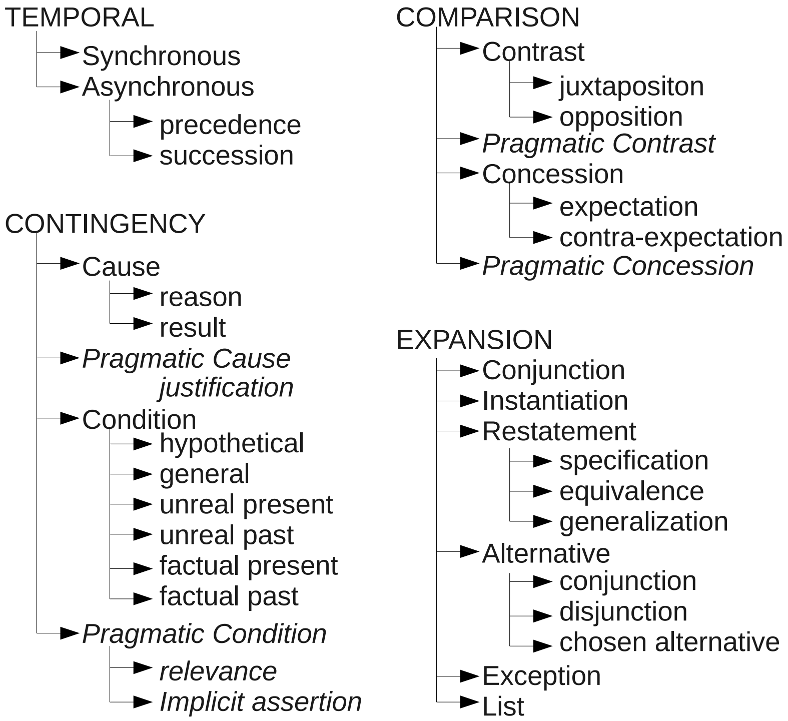


**Figure A.1:** Hierarchy of relation senses in PDTB 2.0 (Prasad et al., 2008).

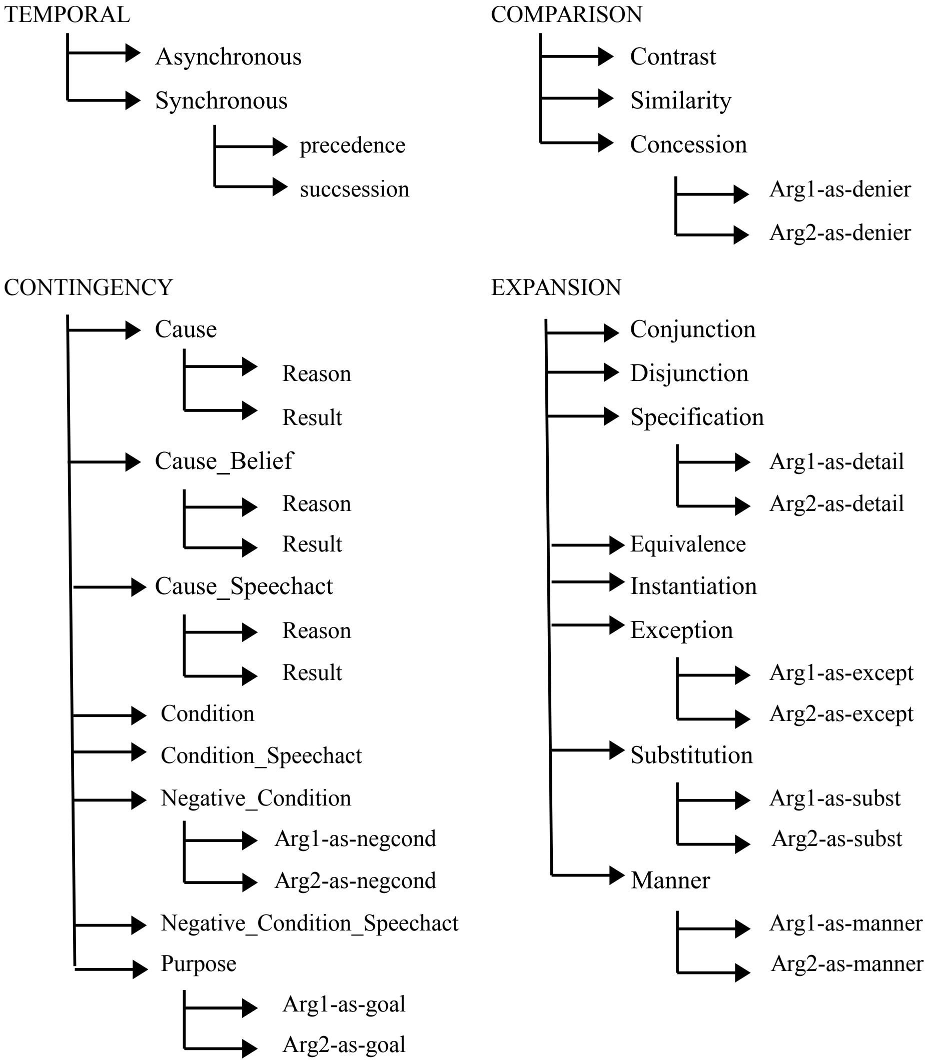
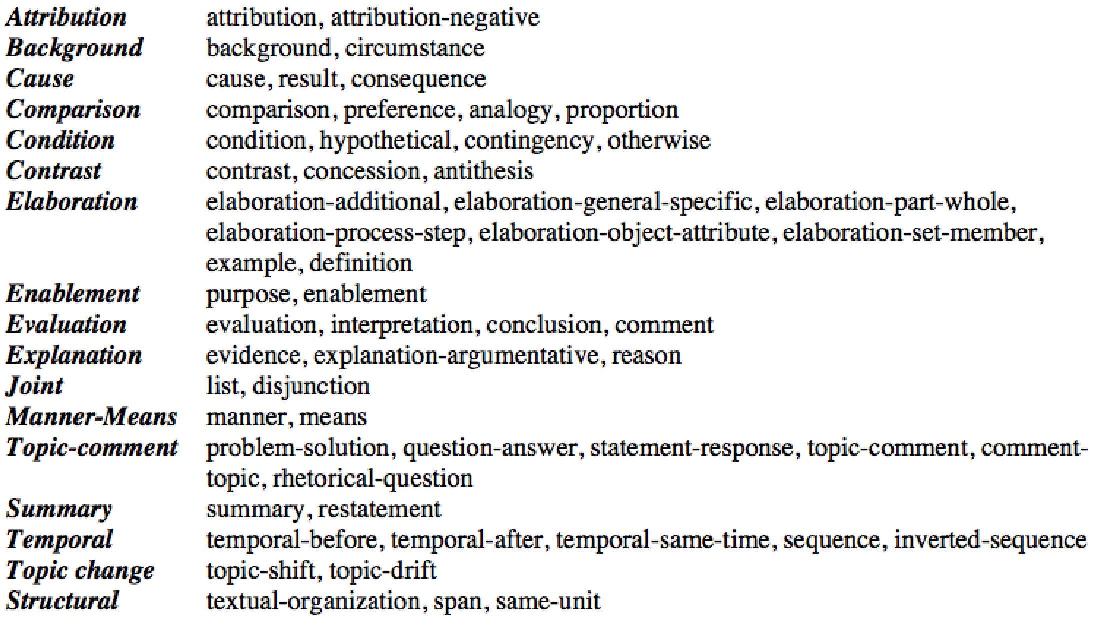


**Figure A.2:** Preliminary hierarchy of relation senses in PDTB 3.0 (see also Rehbein et al., 2016).

# Appendix B

# RST-DT tagset

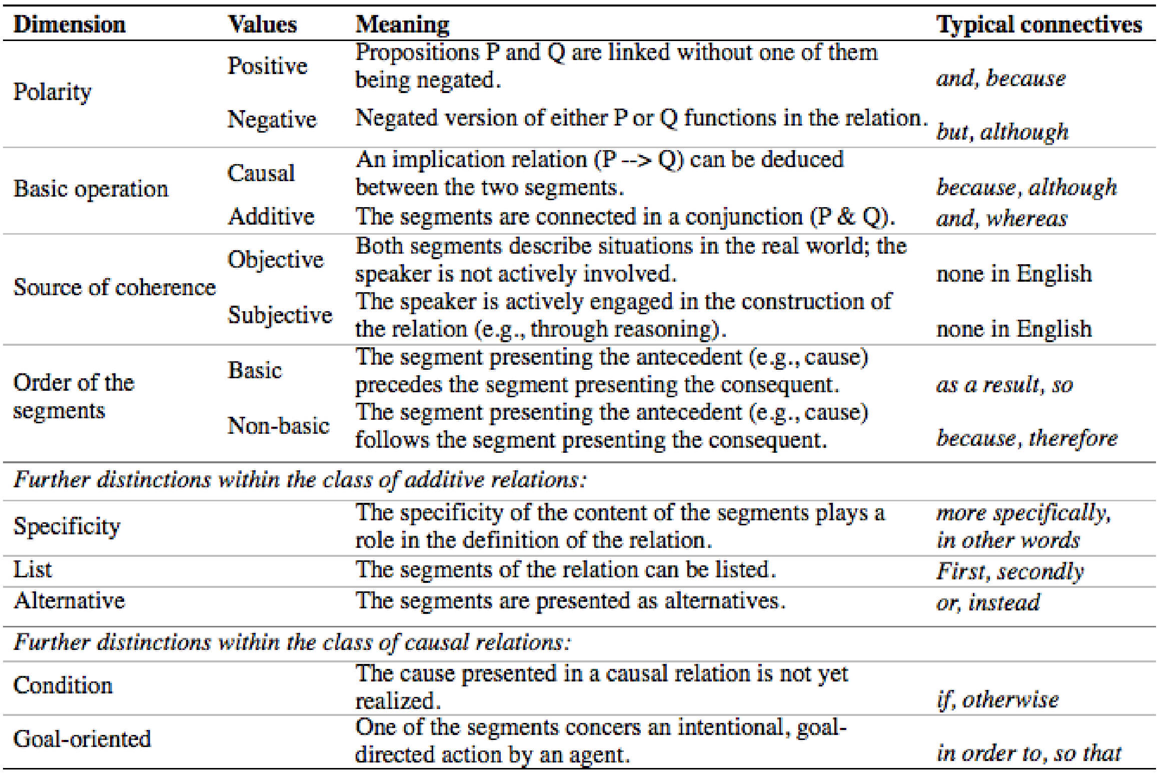


| | |
|---|---|
| ***Attribution*** | attribution, attribution-negative |
| ***Background*** | background, circumstance |
| ***Cause*** | cause, result, consequence |
| ***Comparison*** | comparison, preference, analogy, proportion |
| ***Condition*** | condition, hypothetical, contingency, otherwise |
| ***Contrast*** | contrast, concession, antithesis |
| ***Elaboration*** | elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition |
| ***Enablement*** | purpose, enablement |
| ***Evaluation*** | evaluation, interpretation, conclusion, comment |
| ***Explanation*** | evidence, explanation-argumentative, reason |
| ***Joint*** | list, disjunction |
| ***Manner-Means*** | manner, means |
| ***Topic-comment*** | problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question |
| ***Summary*** | summary, restatement |
| ***Temporal*** | temporal-before, temporal-after, temporal-same-time, sequence, inverted-sequence |
| ***Topic change*** | topic-shift, topic-drift |
| ***Structural*** | textual-organization, span, same-unit |

**Figure B.1:** Tagset of relation senses in RST-DT (Carlson & Marcu, 2001).

# Appendix C

# Unifying Dimensions



| **Dimension** | **Values** | **Meaning** | **Typical connectives** |
|---|---|---|---|
| Polarity | Positive | Propositions P and Q are linked without one of them being negated. | *and, because* |
| | Negative | Negated version of either P or Q functions in the relation. | *but, although* |
| Basic operation | Causal | An implication relation (P --> Q) can be deduced between the two segments. | *because, although* |
| | Additive | The segments are connected in a conjunction (P & Q). | *and, whereas* |
| Source of coherence | Objective | Both segments describe situations in the real world; the speaker is not actively involved. | none in English |
| | Subjective | The speaker is actively engaged in the construction of the relation (e.g., through reasoning). | none in English |
| Order of the segments | Basic | The segment presenting the antecedent (e.g., cause) precedes the segment presenting the consequent. | *as a result, so* |
| | Non-basic | The segment presenting the antecedent (e.g., cause) follows the segment presenting the consequent. | *because, therefore* |
| *Further distinctions within the class of additive relations:* | | | |
| Specificity | | The specificity of the content of the segments plays a role in the definition of the relation. | *more specifically,*<br>*in other words* |
| List | | The segments of the relation can be listed. | *First, secondly* |
| Alternative | | The segments are presented as alternatives. | *or, instead* |
| *Further distinctions within the class of causal relations:* | | | |
| Condition | | The cause presented in a causal relation is not yet realized. | *if, otherwise* |
| Goal-oriented | | One of the segments concers an intentional, goal-directed action by an agent. | *in order to, so that* |

**Figure C.1:** An overview of the Unifying Dimensions and features (see also Sanders et al., 2018).

# Appendix D

# Unifying Dimensions mapping for RST-DT and PDTB

Table D.1, shown on the next page, presents the mapping for RST-DT and PDTB 2.0 labels based on the Unifying Dimensions decomposition.

| PDTB → | TEMP. | | | CONTIN. | | | | | | COMPARISON | | | | | | | EXPANSION | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RST-DT ↓ | Synchrony | precedence | succession | reason | result | justification | Condition | relevance | implicit assertion | Contrast | juxtaposition | opposition | Pragm. contrast | expectation | contra-expectation | Pragm. concession | Conjunction | Instantiation | specification | equivalence | generalization | conjunctive | disjunctive | chosen alternative | Exception | List |
| Background | | x | x | | | | | | | | | | | | | | x | | | | | | | | | |
| Circumstance | | x | x | | | | | | | | | | | | | | x | | | | | | | | | |
| Cause | | | | x | x | | | | | | | | | | | | | | | | | | | | | |
| Cause-result | | | | x | x | | | | | | | | | | | | | | | | | | | | | |
| Result | | | | x | x | | | | | | | | | | | | | | | | | | | | | |
| Consequence | | | | x | x | | | | | | | | | | | | | | | | | | | | | |
| Comparison | | | | | | | | | | | | | | | | | x | | | | | | | | | |
| Preference | | | | | | | | | | x | x | x | x | | | | | | | | | | | x | | |
| Analogy | | | | | | | | | | | | | | | | | x | | | | | | | | | |
| Proportion | | | | | | | x | x | x | | | | | | | | x | | | | | | | | | |
| Condition | | | | | | | x | x | x | | | | | | | | | | | | | | | | | |
| Hypothetical | | | | | | | x | x | x | | | | | | | | | | | | | | | | | |
| Contingency | | | | | | | x | | | | | | | | | | | | | | | | | | | |
| Otherwise | | | | | | | | | | | | | | | | | | | | | | | x | | x | |
| Contrast | | | | | | | | | | x | x | x | x | | | | | | | | | | | | | |
| Concession | | | | | | | | | | | | | | x | x | x | | | | | | | | | | |
| Antithesis | | | | | | | | | | x | x | x | x | x | x | x | | | | | | | | | | |
| El.-additional | | | | | | | | | | | | | | | | | x | | | | | | | | | |
| El.-gen.-spec. | | | | | | | | | | | | | | | | | | | x | x | x | | | | | |
| El.-part-whole | | | | | | | | | | | | | | | | | | | x | x | x | | | | | |
| El.-process-step | | | | | | | | | | | | | | | | | | | x | x | x | | | | | |
| El.-object-attr. | | | | | | | | | | | | | | | | | | | x | x | x | | | | | |
| El.-set-member | | | | | | | | | | | | | | | | | | x | | | | | | | | |
| Example | | | | | | | | | | | | | | | | | | x | | | | | | | | |
| Definition | | | | | | | | | | | | | | | | | | | x | x | x | | | | | |
| Purpose | | | | x | x | x | | | | | | | | | | | | | | | | | | | | |
| Enablement | | | | x | x | x | | | | | | | | | | | | | | | | | | | | |
| Evaluation | | | | x | x | x | | | | | | | | | | | | | x | x | x | | | | | |
| Conclusion | | | | x | x | x | | | | | | | | | | | | | | | | | | | | |
| Comment | | | | | | | | | | | | | | | | | x | | x | x | x | | | | | |
| Evidence | | | | x | x | x | | | | | | | | | | | | | | | | | | | | |
| Expl.-argument. | | | | x | x | | | | | | | | | | | | | | | | | | | | | |
| Reason | | | | x | x | | | | | | | | | | | | | | | | | | | | | |
| List | | | | | | | | | | | | | | | | | | | | | | | | | | x |
| Disjunction | | | | | | | | | | | | | | | | | | | | | | x | x | x | | |
| Summary | | | | | | | | | | | | | | | | | | | | | x | | | | | |
| Restatement | | | | | | | | | | | | | | | | | | | | x | | | | | | |
| Temp.-before | | x | x | | | | | | | | | | | | | | | | | | | | | | | |
| Temp.-after | | x | x | | | | | | | | | | | | | | | | | | | | | | | |
| Temp.-same-time | x | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sequence | | x | x | | | | | | | | | | | | | | | | | | | | | | | |
| Inverted-seq. | | x | x | | | | | | | | | | | | | | | | | | | | | | | |
| Means | | | | x | x | | | | | | | | | | | | | | | | | | | | | |
| Problem-sol.-n | | | | x | x | | | | | | | | | | | | | | | | | | | | | |

**Table D.1:** Proposed mapping between RST-DT and PDTB 2.0 labels according to the Unifying Dimensions approach.

# Appendix E

# Overview of cognitively plausible evidence for the Unifying Dimensions

Table E.1 presents an overview of the evidence regarding the cognitive plausibility of distinctions made in the Unifying Dimensions proposal (see Chapters 3 and 5). The source are presented below the table. Note that the list of sources is not exhaustive; however, it can be assumed that for those cells that remain empty (NA), the literature study revealed no relevant sources.

| **Dimension / feature** | **Evidence from** | | | | | | |
|---|---|---|---|---|---|---|---|
| | Production | Repres./Compr. | Processing | Acquisition | Linguistic system | Annotation | Cross-linguistic |
| Polarity | ✓[1] | ✓[2] | ✓[3] | ✓[4] | ✓[5] | ✓[6] | ✓[7] |
| Basic operation | ✓[8] | ✓[9] | ✓[10] | ✓[11] | ✓[12] | ✓[13] | ✓[14] |
| Source of coherence | ✓[15] | ✓[16] | ✓[17] | ✓[18] | ✓[19] | ✗[20] | ✓[21] |
| Implication order | ✓[22] | NA | ✓[23] | ✓[24], ✗[25] | ✓[26] | ✓[27] | ✓[28] |
| Temporality | ✓[29] | ✓[30] | ✓[31] | ✓[32] | ✓[33] | ✓[34] | ✓[35] |
| Specificity | NA | NA | NA | NA | ✓[36] | ✓[37], ✗[38] | ✓[39] |
| List | NA | NA | ✓[40] | NA | ✓[41] | ✗[42] | ✗[43] |
| Alternative | NA | NA | NA | NA | ✓[44] | ✓[45], ✗[46] | ✓[47] |
| Goal-orientedness | NA | ✓[48] | NA | NA | ✓[49] | ✓[50] | ✓[51] |
| Conditionality | ✓[52] | NA | NA | ✓[53] | ✓[54] | ✓[55] | ✓[56] |

**Table E.1:** An overview of the evidence from various sources regarding the cognitive plausibility of distinctions in the Unifying Dimensions proposal. ✓: evidence; ✗: counter-evidence; NA: not available.

**Sources regarding the cognitive plausibility of Polarity:**

[1] Participants are more likely to continue prompts with a positive relation than a negative relation (see, e.g., results from Kehler et al., 2008).

[2] Recall for positive causal relations is better than for negative causal relations (Townsend, 1983).

[3] Positive causal relations are easier to process than negative causal (i.e., concessive) relations (Drenhaus et al., 2014; Köhne & Demberg, 2013; Xu et al., 2015).

[4] Children produce positive connectives before negative ones (Bloom et al., 1980). Children are more accurate in judging the coherence of positive causal than negative causal relations (Knoepke et al., 2017).

[5] Positive and negative relations have distinct connectives that can mark them, such as *because* and *also* for positive relations, and *however* and *whereas* for negative relations (see, e.g., Prasad et al., 2007).

[6] Annotators can reliably distinguish between positive and negative relations (Scholman et al., 2016).

[7] Children produce positive connectives before negative ones in Dutch (Evers-Vermeul & Sanders, 2009). Positive and negative relations have distinct connectives that can mark them in other languages (e.g., Knott & Sanders, 1998).

**Sources regarding the cognitive plausibility of Basic operation:**

[8] Participants are more likely to continue prompts with a causal relation than a non-causal relation (Kehler et al., 2008).

[9] Causally related sentences are recalled better than non-causally related sentences (Myers et al., 1987; Sanders & Noordman, 2000; Trabasso & Van Den Broek, 1985).

[10] Causally related sentences are easier to process than non-causally related sentences (Black & Bern, 1981; Haberlandt & Bingham, 1978; Kuperberg et al., 2011; Sanders & Noordman, 2000).

[11] Children produce additive connectives before causal connectives in English (Bloom et al., 1980).

[12] Causal and additive relations have distinct connectives and cue phrases that can mark them, such as *because* for causals and *in addition* for additives (see, e.g., Knott & Dale, 1994; Prasad et al., 2007).

[13] Annotators showed high agreement on the annotation of causal and additive relations (Sanders et al., 1992).

[14] Children produce additive connectives before causal ones in Dutch (Evers-Vermeul & Sanders, 2009). Causal and additive relations have distinct connectives that can marked them in various languages (see, e.g., Knott & Sanders, 1998).

**Sources regarding the cognitive plausibility of Source of coherence:**

[15] Objective concessive relations occur more frequently than subjective concessive relations in Dutch (Noordman & Rijswijk, 1997).

[16]Verification times for objective conditional relations are shorter than for subjective relations(Noordman & de Blijzer, 2000).

[17]Objective causal relations are processed faster than subjective causal relations (Canestrelli et al., 2013; Noordman & de Blijzer, 2000; Traxler et al., 1997).

[18]Children produce and comprehend objective causal relations before subjective causal relations (Van Veen, 2011, Ch.5, Ch.6).

[19]In certain languages, subjectivity can be expressed at the level of the connective. For example, Dutch *omdat* ('because) expresses an objective causal meaning, whereas *want* (also 'because') expresses a subjective causal meaning (Stukker & Sanders, 2012).

[20]Annotators often show disagreement on the annotation of objective and subjective relations (see, e.g., Sanders et al., 1992; Scholman et al., 2016).

[21]Children produce both English and German objective causal relations before subjective causal relations (Van Veen, 2011, Ch.5). Various languages have causal connectives that are prototypically used in either subjective or objective relations (Stukker & Sanders, 2012).

**Sources regarding the cognitive plausibility of Implication order:**

[22]Causal relations marked by *because* occur in non-basic order relatively often (but conditional relations do occur frequently in basic order) (Diessel & Hetterle, 2011).

[23]Objective causal relations with a basic order are processed faster than those with a non-basic order (Noordman & de Blijzer, 2000). Causal relations with a non-basic order are easier to process than those with a basic order (Magliano et al., 1993).

[24]Children age 6-7 misinterpret relations with a non-basic order, but not those with a basic order. At age 10 near-perfect comprehension is shown (Bebout et al., 1980).

[25]Children's processing of causal relations was not affected by the implication order of the segments (van den Bosch et al., 2018).

[26]Positive and negative causal relations have corresponding connectives that can express a basic or non-basic order, such as *consequently* versus *because*, and *nevertheless* versus *even though* (Prasad et al., 2007).

[27]Annotators can distinguish between basic and non-basic order of relations relatively well (Sanders et al., 1992; Scholman et al., 2016).

[28]Relations with a basic order are more often implicitated in translation than relations with a non-basic order (Hoek et al., 2017b). Positive and negative causal relations have corresponding connectives that can express a basic or non-basic order in various languages (see, e.g., Knott & Sanders, 1998).

**Sources regarding the cognitive plausibility of Temporality:**

[29]Complex sentences tend to be organized chronologically (Diessel, 2008).

[30]Chronologically ordered sentences facilitate processing (Baker, 1978; Clark & Clark, 1968; Townsend, 1983).

[31]Comprehenders invest more cognitive energy when processing anti-chronological relations (Münte et al., 1998; Ye et al., 2012).

[32]The first emergence of temporal connectives happens after additive connectives, but before causal connectives (Bloom et al., 1980). Children's comprehension of temporal connectives develops in clear stages (Clark, 1971, see also Blything, Davies & Cain, 2015; Pyykkönen & Järvikivi, 2012.

[33]There is ample evidence for connectives and cue phrases that mark temporal relations, such as *before* and *as soon as* (Knott & Dale, 1994, see also Evers-Vermeul, Hoek & Scholman, 2017).

[34]Annotators can distinguish between temporal and non-temporal relations relatively well (Wolf & Gibson, 2005). There is also a relatively large amount of agreement between PDTB and RST-DT on the annotation of temporal relations (see Chapter 4).

[35]Children produce temporal connectives after additive ones but before causal connectives in other languages as well (see, e.g., Evers-Vermeul & Sanders, 2009; van Veen et al., 2014). Many languages have connectives that prototypically mark temporal relations, such as Dutch *toen* and German *danach*.

**Sources regarding the cognitive plausibility of the Specificity feature:**

[36]The relations that can be distinguished within the class of specificity tend to be expressed implicitly in the PDTB (Prasad et al., 2007), but they do have prototypical markers, such as *for example* and *specifically* (Knott, 1996; Prasad et al., 2007).

[37]Annotators can distinguish between Instantiation and Generalization relatively well (Wolf & Gibson, 2005).

[38]There is a relatively large amount of disagreement between PDTB and RST-DT on the annotation of Instantiation and Specification relations (see Chapter 4).

[39]Various languages have markers that express relations within the class of Specificity, such as Spanish *a saber* ('that is to say') and Catalan *és a dir (que)* ('that is').

**Sources regarding the cognitive plausibility of the List feature:**

[40]List relations are processed slower than Problem-solution relation (Sanders & Noordman, 2000). Note that this effect could be attributed to more general differences between additive and causal relations.

[41]List relations can be expressed by the connectives *also* and *moreover*, but these connectives can also express general additive relations in the PDTB (Prasad et al., 2007). Numerical markers such as *first...second* can unambiguously mark List relations.

[42]There is a relatively large amount of disagreement between PDTB and RST-DT on the annotation of LIST relations (see Chapter 4). PDTB 2.0 distinguishes LIST, PDTB 3.0 does not, which speaks against the cognitive plausibility of this distinction.

[43]The LIST distinctions was merged with Conjunction in the Chinese adaptation of the PDTB (Zhou & Xue, 2015).

**Sources regarding the cognitive plausibility of the Alternative feature:**

[44]DISJUNCTIVE and CONJUNCTIVE relations have specific corresponding connectives, namely *alternatively* and *or* (Prasad et al., 2007). CHOSEN ALTERNATIVE relations have a variety of specific markers: the connective *instead*, event modals and negation words (Asr & Demberg, 2015; Knott, 1996; Webber, 2013).

[45]Annotators showed an acceptable amount of agreement on DISJUNCTION relations in the Disco-SPICE corpus (see Chapter 4).

[46]PDTB 2.0 distinguishes DISJUNCTIVE and CONJUNCTIVE; these are merged in PDTB 3.0. This indicates that polarity within the class of alternative relations is not a valid distinction.

[47]Many languages have connectives that prototypically mark DISJUNCTIONS, e.g., Dutch *of*, German *oder*, French *ou* (Hoek, 2018).

**Sources regarding the cognitive plausibility of the Goal-orientedness feature:**

[48]Comprehenders are able to correctly identify PURPOSE event pairs as PURPOSE relations rather than RESULT relations (Andersson & Spenader, 2014).

[49]Relation types belonging to the class of Goal-oriented relations have various prototypical markers. PURPOSE relations can be expressed by *in order to* and/or the modal auxiliaries *can/could* and *will/would* in combination with *so* (Andersson & Spenader, 2014). PROBLEM-SOLUTION relations can be marked by *because* and *therefore*, or lexical signaling devices such as *the solution is* (Sanders & Noordman, 2000). *Means* relations are often preceded by the cue phrase *by* (Carlson & Marcu, 2001).

[50]Annotators showed high agreement on the coding of RESULT versus PURPOSE relations (Andersson & Spenader, 2014).

[51]The PURPOSE distinction was added in the Chinese adaptation of the PDTB (Zhou & Xue, 2015).

**Sources regarding the cognitive plausibility of the Conditionality feature:**

[52]Unlike causal relations, conditional relations tend to be expressed in basic order (Diessel, 2008; Diessel & Hetterle, 2011).

[53]Children produce the conditional connective *if* after they acquire other causal connectives (Bloom et al., 1980).

[54] Condition relations are almost always explicitly marked (Prasad et al., 2007). Prototypical markers include *if* (Dancygier & Sweetser, 2000) and *otherwise* (Webber et al., 1999).

[55] Annotators showed high agreement on the coding of conditional relations (Wolf & Gibson, 2005).

[56] Conditional relations are left explicit in translation more often than other types of relations (Hoek et al., 2017b).

# Appendix F

# Experimental items connective insertion experiments, Chapters 6, 7 and 8

The Wall Street Journal article identifier, inserted PDTB connective, PDTB label and RST-DT label for 114 PDTB Instantiation and Specification items used in the connective insertion experiments in Chapters 6 and 7 are presented below. The WSJ IDs of items that were also included in Experiment 1 in Chapter 8 are printed in bold; those that were included in Experiment 2 in Chapter 8 are underlined.

The identifiers, rather than the items, are presented here due to copyright restrictions. The corresponding relations can be found in the Penn Discourse Treebank (Prasad et al., 2007) or the Rhetorical Structure Theory Discourse Treebank (Carlson et al., 2003).

**WSJ ID: 601**; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.
WSJ ID: 610; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.
WSJ ID: 616; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: elaboration-general-specific.
WSJ ID: 617; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.
WSJ ID: 629; PDTB connective: indeed; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.
WSJ ID: 629; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.
WSJ ID: 629; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

**WSJ ID: 629**; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.
WSJ ID: 634; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.
WSJ ID: 640; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.
WSJ ID: 640; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.
**WSJ ID: 664**; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: elaboration-general-specific.
WSJ ID: 666; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.
WSJ ID: 671; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.
**WSJ ID: 671**; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.
**WSJ ID: 671**; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.
**WSJ ID: 675**; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.
WSJ ID: 676; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.
WSJ ID: 677; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.
WSJ ID: 677; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.
WSJ ID: 681; PDTB connective: in other words; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.
**WSJ ID: 681**; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: elaboration-general-specific.
WSJ ID: 687; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.
WSJ ID: 687; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.
WSJ ID: 690; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: elaboration-general-specific.
WSJ ID: 1105; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.
WSJ ID: 1120; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.
**WSJ ID: 1120**; PDTB connective: for example; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1121; PDTB connective: for example; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

**WSJ ID: 1128**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

**WSJ ID: 1128**; PDTB connective: indeed; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1136; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

WSJ ID: 1137; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

**WSJ ID: 1137**; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

WSJ ID: 1140; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

**WSJ ID: 1141**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1145; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

WSJ ID: 1146; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

**WSJ ID: 1151**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

**WSJ ID: 1154**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

**WSJ ID: 1154**; PDTB connective: for instance; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

WSJ ID: 1160; PDTB connective: for instance; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

**WSJ ID: 1162**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1162; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: elaboration-general-specific.

WSJ ID: 1162; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1162; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

**WSJ ID: 1162**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

WSJ ID: 1163; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

WSJ ID: 1166; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

WSJ ID: 1172; PDTB connective: for instance; PDTB label: Expansion.Instantiation; RST-DT label: elaboration-general-specific.

WSJ ID: 1172; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 1174; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

WSJ ID: 1174; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

WSJ ID: 1174; PDTB connective: for instance; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 1174; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 1187; PDTB connective: for instance; PDTB label: Expansion.Instantiation; RST-DT label: elaboration-general-specific.

WSJ ID: 1192; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 1192; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 1192; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

**WSJ ID: 1302**; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

**WSJ ID: 1311**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1314; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1314; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

WSJ ID: 1315; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

**WSJ ID: 1317**; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1317; PDTB connective: for example; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

WSJ ID: 1318; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

**WSJ ID: 1319**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

WSJ ID: 1319; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

**WSJ ID: 1320**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

**WSJ ID: 1320**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1320; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1330; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1331; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

WSJ ID: 1332; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

WSJ ID: 1337; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

WSJ ID: 1366; PDTB connective: for instance; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

WSJ ID: 1367; PDTB connective: for instance; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

WSJ ID: 1368; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 1368; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

WSJ ID: 1368; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 1373; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

WSJ ID: 1377; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

WSJ ID: 1380; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

WSJ ID: 1390; PDTB connective: for instance; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 1394; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

WSJ ID: 1394; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

WSJ ID: 1394; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: elaboration-general-specific.

WSJ ID: 1394; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: elaboration-general-specific.

WSJ ID: 1394; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

WSJ ID: 1394; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: evidence.

WSJ ID: 1924; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

WSJ ID: 1962; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

**WSJ ID: 1970**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

WSJ ID: 1984; PDTB connective: for instance; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

WSJ ID: 2309; PDTB connective: in fact; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

WSJ ID: 2321; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

WSJ ID: 2325; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: evidence.

**WSJ ID: 2325**; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

WSJ ID: 2331; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: elaboration-general-specific.

WSJ ID: 2341; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

WSJ ID: 2343; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 2345; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

WSJ ID: 2346; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

WSJ ID: 2346; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: example.

WSJ ID: 2350; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

WSJ ID: 2366; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

WSJ ID: 2366; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

WSJ ID: 2366; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

WSJ ID: 2381; PDTB connective: for example; PDTB label: Expansion.Instantiation; RST-DT label: explanation-argumentative.

WSJ ID: 2391; PDTB connective: specifically; PDTB label: Expansion.Restatement.Specification; RST-DT label: example.

WSJ ID: 2398; PDTB connective: in particular; PDTB label: Expansion.Restatement.Specification; RST-DT label: explanation-argumentative.

# Appendix G

# Experimental items working memory study, Chapter 8

The 56 experimental sentences used in the working memory study are presented below. Half of the sentences contain a verb that requires an animate subject and contain a verb that requires an animate objject. Sentences belonged to one of four possible types: cleft subject, cleft object, object-subject and subject-object. Finally, half of the sentences are acceptable and half are unacceptable.

**1.** It was the elephant that escaped from the zoo. *(Animate subj., cleft subj., acceptable.)*
**2.** It was the professor that forgot the handout. *(Animate subj., cleft subj., acceptable.)*
**3.** It was the man that clenched the pillow. *(Animate subj., cleft subj., acceptable.)*
**4.** It was the bear that made the growling sound. *(Animate subj., cleft subj., acceptable.)*
**5.** It was the sound that made the computer. *(Animate subj., cleft subj., unacceptable.)*
**6.** It was the document that filed the librarian. *(Animate subj., cleft subj., unacceptable.)*
**7.** It was the cookie that ate the talented dancer. *(Animate subj., cleft subj., unacceptable.)*
**8.** It was the toy that fascinated the child. *(Animate obj., cleft subj., acceptable.)*
**9.** It was the porcelain doll that scared the little girl. *(Animate obj., cleft subj., acceptable.)*
**10.** It was the bike that hit the news reporter. *(Animate obj., cleft subj., acceptable.)*
**11.** It was the building that impressed the architect. *(Animate obj., cleft subj., acceptable.)*
**12.** It was the sulky teenager that bored the book. *(Animate obj., cleft subj., unacceptable.)*
**13.** It was the florist that astonished the bouquet. *(Animate obj., cleft subj., unacceptable.)*
**14.** It was the family that shocked the revelation. *(Animate obj., cleft subj., unacceptable.)*
**15.** It was the teddy bear that the child wanted. *(Animate subj., cleft obj., acceptable.)*
**16.** It was the Polish bread that the family liked to eat. *(Animate subj., cleft obj., acceptable.)*
**17.** It was the upbeat pop song that the grandfather hated. *(Animate subj., cleft obj., acceptable.)*
**18.** It was the young man that the engagement ring bought. *(Animate subj., cleft obj.,*

*unacceptable.)*

**19.** It was the CEO that the enormous file requested. *(Animate subj., cleft obj., unacceptable.)*

**20.** It was the owner of the house that the light fixed. *(Animate subj., cleft obj., unacceptable.)*

**21.** It was the model that the cheeseburger ate. *(Animate subj., cleft obj., unacceptable.)*

**22.** It was the older man that the speech offended. *(Animate obj., cleft obj., acceptable.)*

**23.** It was the widow that the vivid dream tormented. *(Animate obj., cleft obj., acceptable.)*

**24.** It was the innocent people that the restrictions affected. *(Animate obj., cleft obj., acceptable.)*

**25.** It was the berry that the amateur hunter poisoned. *(Animate obj., cleft obj., unacceptable.)*

**26.** It was the memory that the heartbroken girl haunted. *(Animate obj., cleft obj., unacceptable.)*

**27.** It was the embroidery that the young girl calmed down. *(Animate obj., cleft obj., unacceptable.)*

**28.** It was the amulet that the innocent children hypnotized. *(Animate obj., cleft obj., unacceptable.)*

**29.** The boy envied the friend that bought a new game. *(Animate subj., obj.-subj., acceptable.)*

**30.** The girl played with the friend that injured her foot. *(Animate subj., obj.-subj., acceptable.)*

**31.** The athlete hired the manager that applied for the position. *(Animate subj., obj.-subj., acceptable.)*

**32.** The scarf loved the designer that kept the model warm. *(Animate subj., obj.-subj., unacceptable.)*

**33.** The milk drank the lawyer that turned sour. *(Animate subj., obj.-subj., unacceptable.)*

**34.** The castle hated the man that interested all tourists. *(Animate subj., obj.-subj., unacceptable.)*

**35.** The sound listened to the man that came from the basement. *(Animate subj., obj.-subj., unacceptable.)*

**36.** The artwork fascinated the girl that loved to paint. *(Animate obj., obj.-subj., acceptable.)*

**37.** The drug worried the pharmacist that worked with patients. *(Animate obj., obj.-subj., acceptable.)*

**38.** The book inspired the young girl that admired strong women. *(Animate obj., obj.-subj., acceptable.)*

**39.** The flower field charmed the lady that had always lived in the city. *(Animate obj., obj.-subj., acceptable.)*

**40.** The woman bothered the rain that had straightened her hair. *(Animate obj., obj.-subj., unacceptable.)*

**4.1** The police officer puzzled the evidence that investigated the case. *(Animate obj., obj.-*

*subj., unacceptable.)*

**42.** The gardener pleased the good weather that worked outside. *(Animate obj., obj.-subj., unacceptable.)*

**43.** The employee that the boss fired yelled at her supervisor. *(Animate subj., subj.-obj., acceptable.)*

**44.** The singer that everybody loved hated rock music. *(Animate subj., subj.-obj., acceptable.)*

**45.** The actress that the audience booed at called her manager. *(Animate subj., subj.-obj., acceptable.)*

**46.** The man that the painter loved despised his parents. *(Animate subj., subj.-obj., acceptable.)*

**47.** The door that nobody trusted pushed the doctor open. *(Animate subj., subj.-obj., unacceptable.)*

**48.** The secrets that the woman was seeing betrayed her psychologist. *(Animate subj., subj.-obj., unacceptable.)*

**49.** The knee that the coach trained injured his gymnast. *(Animate subj., subj.-obj., unacceptable.)*

**50.** The meat that the butcher cut delighted the customer. *(Animate obj., subj.-obj., acceptable.)*

**51.** The pen that the teacher brought splashed ink on the children. *(Animate obj., subj.-obj., acceptable.)*

**52.** The song that the mother played comforted the sad daughter. *(Animate obj., subj.-obj., acceptable.)*

**53.** The man that the doctor recommended helped the exercises. *(Animate obj., subj.-obj., unacceptable.)*

**54.** The man that the plumber didn't fix annoyed the leaking shower. *(Animate obj., subj.-obj., unacceptable.)*

**55.** The customers that the woman made impressed the bracelet. *(Animate obj., subj.-obj., unacceptable.)*

**56.** The audience that the acrobats performed astonished the trick. *(Animate obj., subj.-obj., unacceptable.)*

# Appendix H

# Corpus example of OT1H and OTOH

The following example passage is taken from the ukWaC corpus.

(50) **On the one hand** there was no group of Yugoslavs to whom the phrasing of Robertson's order applied more completely than the Croat Ustachi, Domobranci and regular troops under German Army Group E, who at the time were approaching the Austrian border, armed and in very large numbers. **Certainly** Gen McCreery wanted specific authorization from AFHQ to deal with these Croats in precisely the terms which Robertson's order gave him. **But** we have no firm indication that, at the moment Robertson drafted the order, he was specifically aware of McCreery's request for such an authorization. **Nor** did Gen Robertson word his order to make it refer exclusively to Croats. **If** he had meant to, he surely would just have stated ”Croats”. That he did not so do indicates he had some other categories in mind as well. **On the other hand**, it is **also** clear that Robertson originally intended to word his order in such a way that some dissident Yugoslavs should be excluded from the hand-overs, **and** these he categorized as “Chetniks”.

In order to understand example (50), the reader must build a discourse structure that accommodates a contrastive relation based on OT1H. Before encountering the second argument of this relation, however, the reader has to process additional explicit discourse markers, namely *certainly*, *but*, *nor*, and *if*.

# Appendix I

# Experimental items Chapter 9

The 24 experimental items used in the norming study, story continuation study, and eye-tracking study reported in Chapter 9 are presented below. Sentence version (i) is a globally contrastive intervening sentence, version (ii) is a globally contrastive intervening sentence, and (iii) is a non-contrastive intervening sentence.

**1.** Michael heard that his favourite singer Beyoncé is coming to Edinburgh during her tour. On the one hand, he's thinking about taking two days off for the concert, because she'll only be in Edinburgh for one show. **(i)** But he might only be able to get one day off work. **(ii)** But she will probably be back next year for more concerts. **(iii)** Also, he still needs to finish some of his vacation days. On the other hand, he has a deadline coming up and really needs to get his work done.

**2.** On a rainy day, Gillian was thinking of asking her friend Mark to join her for a shopping trip. On the one hand, she was thinking that they could take the car, because he just passed his driver's license exam last month. **(i)** But she also knows that the bus ride is pretty quick. **(ii)** But he might not feel comfortable driving in the rain. **(iii)** Also, they will be able to play their own music in the car. On the other hand, she might get everything she needs faster if she just goes shopping by herself.

**3.** Jon is from Spain and is considering going to a Scottish ceilidh, to dance and listen to music. On the one hand, he thinks it might be a lot of fun, because he's heard great stories about these parties from his brother. **(i)** But he doesn't know anybody else who will be there. **(ii)** But he does not have the same taste in music as his brother. **(iii)** Also, he would like to learn more about Scottish culture. On the other hand, he's really worried about other people seeing his underwear when dancing with a kilt.

**4.** Mary is thinking about taking part in the whisky tasting at the Talisker distillery in Scotland. On the one hand, she's really curious about trying real Scottish whisky, because she's read a lot about the smoky smells. **(i)** But she is a lightweight and doesn't like getting drunk. **(ii)** But she has a blocked nose and doesn't smell much. **(iii)** Also, she has heard that it's different from American whisky. On the other hand, she feels like the whisky tasting could be too expensive for her travel budget.

**5.** Peter is looking for Scottish recipes with which he can impress his visitors from overseas. On the one hand, haggis would be a good dish to serve, because he used to love his mother's haggis when he was a child. **(i)** But not everyone likes sheep intestines and brains. **(ii)** But his mother is actually a much better cook than he is. **(iii)** Also, it is something that's very unique to Scotland. On the other hand, he won't be able to find the special utensils to prepare the haggis anyway.

**6.** John has been dating Sue for a few months and he's thinking about their future together. On the one hand, he'd like to buy a bigger house and move in with her right away, because she can cook amazingly well. **(i)** But he would like to enjoy the bachelor life a little longer. **(ii)** But she is terrible at remembering to wash the dishes. **(iii)** Also, she is great at making a house feel like a cozy home. On the other hand, he might just take things slow and give her the keys to his apartment.

**7.** Bob suggested a business merger with Jennifer's company, and now she's considering it. On the one hand, she'd like to join forces with Bob, because he already has many loyal, and even some famous customers. **(i)** But she would rather avoid the costly legal fees of a merger. **(ii)** But she is worried the celebrities will have high demands. **(iii)** Also, he is known for having a lot of business experience. On the other hand, she wants to make sure she can rise to power as CEO without competition.

**8.** Susan doesn't like her job at the warehouse and is mulling over what to do next with life. On the one hand, she might start a farm somewhere in New Zealand, because New Zealand is the sheep capital of the world. **(i)** But she could open a scuba diving school in Australia. **(ii)** But the wool industry has been going downhill these days. **(iii)** Also, she has heard the nature is beautiful in New Zealand. On the other hand, she could also move to India and join an Ashram in order to meditate.

**9.** Joseph got a job offer from the Edinburgh Zoo and he's pondering whether he should take it. On the one hand, he needs the money that this job will pay, because he should start paying off his student loans this year. **(i)** But he could keep looking for a nicer, better-paying job. **(ii)** But the loans could be deferred for a few more months. **(iii)** Also, his car needs to be serviced by the end of the month. On the other hand, he hates the idea of cleaning out panda cages and lions' dens every day.

**10.** Maryann is considering taking surfing lessons during her vacation in Hawaii next month. On the one hand, she loves the idea of surfing in Hawaii, because she's heard that there are many cute surfer boys there. **(i)** But she is afraid her painful back will not be up to it. **(ii)** But she already has a boyfriend whom she loves a lot. **(iii)** Also, she has heard that the waves are really high in Hawaii. On the other hand, relaxing at the beach with a nice cocktail sounds very good to her too.

**11.** Kate wants to go visit her brother in Aberdeen with her newborn baby during the weekend. On the one hand, she's thinking about driving there, because it's less of a hassle than traveling by train with a newborn baby. **(i)** But the traffic in Aberdeen is always busy and chaotic. **(ii)** But her baby has always been quiet on trains so far. **(iii)** Also, going by car will probably be faster than by train. On the other hand, she's thinking the baby might be too young to spend so much time traveling.

**12.** Frank is thinking about quitting his job at the supermarket after working there for five years. On the one hand, he thinks he could get a more promising job at a multinational, because he studied accounting in college. **(i)** But he would miss the personal contact with customers. **(ii)** But he has no real work experience as an accountant. **(iii)** Also, he knows someone who could get him an interview. On the other hand, he has a good chance at becoming a manager at the supermarket next year.

**13.** Lisa found out that she is unexpectedly pregnant and is unsure what to do with the baby. On the one hand, she's thinking that she'd like to keep the baby, because she always loves playing with her baby nephew. **(i)** But she is not sure whether her boyfriend wants a baby. **(ii)** But she usually sees her nephew for a few hours only. **(iii)** Also, she has always dreamed about being a good mother. On the other hand, she always wanted to be married and have a house before having a baby.

**14.** Daniel has been dating a girl from work and he's considering introducing her to his parents. On the one hand, he's sure his mother will be excited, because she thinks that, at his age, he should be married already. **(i)** But he thinks his mother will dislike his girlfriend's tattoos. **(ii)** But he is unsure whether he'd want to marry this girl. **(iii)** Also, she is always curious about the girls he's seeing. On the other hand, his girlfriend might think it's too soon to meet his parents and get scared.

**15.** Nicole is turning 27 next week and she's mulling over what snacks to serve at her party. On the one hand, she'd like to prepare finger food, because she wants to show her mother that she has improved her cooking. **(i)** But she wants to spend very little time in the kitchen. **(ii)** But she is unsure if her mother will come to the party. **(iii)** Also, she thinks it's classy to serve finger food at a party. On the other hand, she can make it easy for herself and only serve some pretzels and nuts.

**16.** Henry's laptop is quite old and now he's debating whether or not he should get a new one. On the one hand, he can afford to buy one now, because he just heard from his boss that he'll receive a bonus this Christmas. **(i)** But the laptop that he has is actually still working fine. **(ii)** But he wanted to spend his bonus on a vacation to Bali. **(iii)** Also, he has some money saved up from his birthday. On the other hand, he would also like a tablet and could use that to check emails too.

**17.** Anthony woke up with a headache and now he's thinking about calling in sick for work today. On the one hand, he won't miss a lot, because he was only planning on attending a talk and had no other meetings planned. **(i)** But he would like to get ahead on his quarterly reports. **(ii)** But he was excited about learning more from the talk. **(iii)** Also, he can check his email and answer calls from home. On the other hand, he might feel a lot better already after taking an aspirin and some vitamins.

**18.** Johanna got an invitation from her aunt to visit her for two weeks in Tanzania this winter. On the one hand, she thinks it could be a great experience, because she would be able to go on a safari for the first time. **(i)** But she fears the African heat she's heard so much about. **(ii)** But she is very afraid of wild animals, especially lions. **(iii)** Also, she is curious about African culture and customs. On the other hand, she's not sure whether she can find someone to take care of her dog.

**19.** Melissa's friend lives at the seaside and now Melissa is planning her weekend trip there. On the one hand, she'd like to drive there directly after work on Friday, because her friends will have a party there that night. **(i)** But the roads to the seaside would be incredibly busy. **(ii)** But someone she doesn't like might also attend. **(iii)** Also, she wants to spend as much time there as possible. On the other hand, relaxing at home on Friday evening would make her feel less stressed and rushed.

**20.** Nan is unsure of what she wants to do after she gets her Bachelor's degree in informatics. On the one hand, she might do a Master's at the same uni, because that'll make it easier to get a research position there. **(i)** But she would like to study abroad once in her life as well. **(ii)** But the research positions at her university are not well paid. **(iii)** Also, it will look good on her CV if she has a Master's degree. On the other hand, she could do a traineeship at Shell and learn more about the corporate world.

**21.** David is thinking about taking his girlfriend out somewhere to improve their relationship. On the one hand, he'd like to invite her to a rock music festival, because his old friends will be playing a short set there. **(i)** But he is not sure whether she really likes rock music. **(ii)** But they would spend a lot of time with his friends then. **(iii)** Also, he likes the other bands that will play at the festival. On the other hand, she might be happier if they watch a romantic movie together at the cinema.

**22.** Gary's favorite holiday is Christmas and now he's wondering how to celebrate it this year. On the one hand, he'd like to go to South Africa, because his parents recently moved there and he'd like to surprise them. **(i)** But he would rather not celebrate Christmas in warm weather. **(ii)** But he has been quarrelling a lot with his father lately. **(iii)** Also, he would like to experience the South African way of life. On the other hand, he heard that the public security of South Africa has a really bad reputation.

**23.** Lucy has a lot of savings and she's thinking about how to manage her personal finances. On the one hand, she'd like to invest in stocks, because her sister has experience in stocks investment and can give her advice. **(i)** But stocks investment is often accompanied by high risk. **(ii)** But her sister has actually made little profit from it so far. **(iii)** Also, she has been told that stocks can be very profitable. On the other hand, she could help out her best friend and invest her money in his business.

**24.** Helen found some signs that Ben, her colleague and friend, has violated the company rules. On the one hand, she's thinking she should report it to a superior, because the violation may lead to losses for the company. **(i)** But this will certainly have an effect on their friendship. **(ii)** But he may stop before he actually causes the losses. **(iii)** Also, she will be an accomplice if she doesn't say anything. On the other hand, she could talk to Ben about it first before reporting it to their superior.

## Appendix J

# Occurrence of connectives in Experiment 2, Chapter 9

Table J.1, presented on the next page, provides information regarding the distribution of connectives in the story continuation study (Experiment 2), reported in Chapter 9.

| **Continuation type** | **Connective** | **Global** | **Local** | **No-contrast** | **Total** |
|---|---|---|---|---|---|
| +contrastive OTOH present | on the other hand | 11 | 31 | 128 | 170 |
| | but on the other hand | 1 | 0 | 5 | 6 |
| | on the other | 0 | 0 | 5 | 5 |
| | and on the other hand | 1 | 2 | 0 | 3 |
| +contrastive OTOH absent | however | 2 | 3 | 25 | 30 |
| | but | 2 | 6 | 10 | 18 |
| | although/though | 0 | 5 | 6 | 11 |
| | alternatively | 1 | 3 | 0 | 4 |
| | also | 0 | 1 | 1 | 2 |
| | on the downside | 0 | 0 | 1 | 1 |
| | otherwise | 0 | 0 | 1 | 1 |
| | conversely | 0 | 0 | 1 | 1 |
| | *(no connective)* | 6 | 33 | 9 | 48 |
| –contrastive | so | 10 | 5 | 3 | 18 |
| | also | 4 | 9 | 2 | 15 |
| | but | 1 | 5 | 0 | 6 |
| | therefore | 0 | 3 | 2 | 5 |
| | however | 2 | 1 | 0 | 3 |
| | although/though | 1 | 0 | 1 | 2 |
| | then | 2 | 0 | 0 | 2 |
| | as a result | 1 | 0 | 0 | 1 |
| | *(no connective)* | 195 | 133 | 40 | 368 |
| Total | | 240 | 240 | 240 | 720 |

**Table J.1:** +contrastive continuations convey content that contrasts with the OT1H-clause. –contrastive continuations convey non-contrastive content or content that contrasts with material outside the OT1H-clause.

# Appendix K

# Experimental items English experiment, Chapter 10

The 16 experimental items used in the English study reported in Chapter 10 are presented below.

**1.** Michael heard that his favourite singer Shakira is coming to Edinburgh during her tour. On the one hand, he's thinking about taking two days off for the concert, because she'll only be in Edinburgh for one show. On the other hand, he has a deadline coming up and really needs to get his work done.
**2.** Heather is considering taking surfing lessons during her vacation in Hawaii next month. On the one hand, she loves the idea of surfing in Hawaii, because she's heard that there are many cute surfer boys there. On the other hand, relaxing at the beach with a nice cocktail sounds very good to her too.
**3.** Jess wants to go visit her brother in Edinburgh with her newborn baby during the weekend. On the one hand, she's thinking about driving there, because it's easier than travelling by train with a newborn baby. On the other hand, she's thinking the baby might be too young to spend so much time travelling.
**4.** Ronan is thinking about quitting his job at the supermarket after working there for five years. On the one hand, he thinks he could get a more promising job at a multinational, because he studied accounting in college. On the other hand, he has a good chance of becoming a manager at the supermarket next year.
**5.** Lisa found out that she is unexpectedly pregnant and is unsure what to do with the baby. On the one hand, she's thinking that she'd like to keep the baby, because she always loves playing with her baby nephew. On the other hand, she always wanted to be married and have a house before having a baby.
**6.** Gordon has been dating a girl from work and he's considering introducing her to his parents. On the one hand, he's sure his mother will be excited, because she thinks that, at his age, he should be married already. On the other hand, his girlfriend might think it's too

soon to meet his parents and get scared.

**7.** Anthony woke up with a headache and now he's thinking about calling in sick for work today. On the one hand, he won't miss a lot, because he was only planning on attending a talk and had no other meetings planned. On the other hand, he might feel a lot better already after taking an aspirin.

**8.** Melissa's friend lives at the seaside and now Melissa is planning her weekend trip there. On the one hand, she'd like to drive there after work on Friday, because her friends will have a party there that night. On the other hand, relaxing at home on Friday evening would make her feel less stressed and rushed.

**9.** On a rainy day, Gillian was thinking of asking her friend Mark to join her for a shopping trip. On the one hand, she was thinking that they could take the car, because he just passed his driver's license exam. On the other hand, she might get everything she needs faster if she just goes shopping by herself.

**10.** Nan is unsure of what she wants to do after she gets her Bachelor's degree in Finance. On the one hand, she might do a Master's at the same uni, because that'll make it easier to get a research position there. On the other hand, she could do a traineeship at Shell and learn more about the corporate world.

**11.** Lucy has a lot of savings and she's thinking about how to manage her personal finances. On the one hand, she'd like to invest in stocks, because her sister has experience in investment and can give her advice. On the other hand, she could invest her money in her friends' business.

**12.** Helen found some signs that Ben, her colleague and friend, has violated the company rules. On the one hand, she's thinking she should report it to a superior, because the violation may lead to losses for the company. On the other hand, she could talk to Ben about it first before reporting it to their superior.

**13.** Jon is from Spain and is considering going to a Scottish ceilidh, to dance and listen to music. On the one hand, he thinks it might be a lot of fun, because he's heard great stories about these parties. On the other hand, he's really worried about other people seeing his underwear when dancing with a kilt.

**14.** Peter is looking for Scottish recipes with which he can impress his visitors from overseas. On the one hand, haggis would be a good dish to serve, because he used to love his mother's haggis when he was a child. On the other hand, he won't be able to find the special utensils to prepare the haggis anyway.

**15.** Bob suggested a business merger with Jennifer's company, and now she's considering it. On the one hand, she'd like to join forces with Bob, because he already has many loyal, and even some famous customers. On the other hand, she wants to make sure she can rise to power as CEO without competition.

**16.** Emily doesn't like her job at the warehouse and is mulling over what to do next with life. On the one hand, she might start a farm somewhere in New Zealand, because New Zealand is the sheep capital of the world. On the other hand, she could also move to India and join an Ashram in order to meditate.

# Appendix L

# Experimental items Dutch experiment, Chapter 10

The 12 experimental items used in the Dutch study reported in Chapter 10 are presented below.

**1.** De werkzoekende overwoog om te solliciteren op een baan in de dierentuin. Aan de ene kant had hij het geld hard nodig. Aan de andere kant leek het hem niet leuk om de kooien schoon te moeten maken.
**2.** De backpackster was haar trip naar Hawaii aan het plannen. Aan de ene kant leek het haar leuk om daar te leren surfen. Aan de andere kant leek het haar ook leuk om te snorkelen in de oceaan.
"**3.** De vakkenvuller overwoog laatst om zijn baan op te zeggen. Aan de ene kant dacht hij dat hij een baan kon krijgen bij een multinational. Aan de andere kant was er een kans dat hij volgend jaar manager werd bij de supermarkt.
" **4.** De vriendin was een verrassingsfeest voor haar vriend aan het plannen. Aan de ene kant wilde ze graag warme hapjes serveren. Aan de andere kant was het makkelijker om nootjes en chips serveren.
**5.** De professor werd wakker met hoofdpijn en wilde zich misschien ziek melden. Aan de ene kant zou hij niet veel missen. Aan de andere kant voelde hij zich waarschijnlijk een stuk beter als hij een paracetamol nam.
**6.** De studente werd door haar oom uitgenodigd om op bezoek te komen in Tanzania. Aan de ene kant dacht ze dat het een fantastische ervaring kon zijn. Aan de andere kant wist ze niet of iemand op haar hond kon passen.
**7.** De studente Informatica wist niet zeker wat ze wilde doen na haar studie. Aan de ene kant kon ze een master doen aan dezelfde universiteit. Aan de andere kant kon ze ook een traineeship doen bij de gemeente.
**8.** De jongeman wilde zijn vriendin dit weekend meenemen op een leuke date. Aan de ene kant wilde hij haar meenemen naar een festival. Aan de andere kant vond zij het misschien

leuker om naar de bioscoop te gaan.

**9.** De ondernemer overwoog om zijn zaak te fuseren met een andere zaak. Aan de ene kant kon hij profiteren van de klantenbasis van de andere ondernemer. Aan de andere kant wilde hij graag als enige de leiding hebben in zijn zaak.

**10.** De avonturierster dacht vorige week na over een nieuw avontuur. Aan de ene kant leek het haar leuk om op een boerderij in Schotland te gaan werken. Aan de andere kant leek het haar ook leuk om te mediteren in India.

**11.** De winnaar van de postcodeloterij dacht na over wat hij kon doen met het geld. Aan de ene kant kon hij zijn geld beleggen. Aan de andere kant kon hij ook zijn vriend helpen door in zijn bedrijf te investeren.

**12.** De serveerster betrapte haar collega erop toen hij uit de fooienpot steelde. Aan de ene kant kon ze hem aangeven bij haar baas. Aan de andere kant kon ze ook eerst met haar collega erover praten.

# List of Figures

# List of Tables

# Bibliography

Aaronson, D., & Ferres, S. (1986). Reading strategies for children and adults: A quantitative model. *Psychological Review*, *93*, 89–112.

Afflerbach, P. (2015). *Handbook of individual differences in reading: Reader, text, and context*. Routledge.

Al-Saif, A., & Markert, K. (2010). The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)* (pp. 2046–2053). Valletta, Malta.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264.

Ambridge, B., & Rowland, C. F. (2013). Experimental methods in studying child language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 149–168.

Andersson, M., & Spenader, J. (2014). Result and Purpose relations with and without 'so'. *Lingua*, *148*, 1–27.

Arai, M., & Keller, F. (2013). The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, *28*, 525–560.

Arnold, J. E. (1998). *Reference form and discourse patterns*. Ph.D. thesis Stanford University.

Artstein, R., & Poesio, M. (2005). Bias decreases in proportion to the number of annotators. *Proceedings of the Conference on Formal Grammar and Mathematics of Language (FG-MoL)*, (pp. 141–150).

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*, 555–596.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer, Norwell, MA.

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

Asr, F. T., & Demberg, V. (2012). Implicitness of discourse relations. In *Proceedings of the International Conference on Computational Linguistics (COLING)* (pp. 2669–2684). Mumbai, India.

Asr, F. T., & Demberg, V. (2013). On the information conveyed by discourse markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics* (pp. 84–93). Sofia, Bulgaria.

Asr, F. T., & Demberg, V. (2015). Uniform Information Density at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the International Conference on Computational Semantics (IWCS)* (pp. 118–128). London, UK.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Baker, L. (1978). Processing temporal relationships in simple stories: Effects of input sequence. *Center for the Study of Reading Technical Report; no. 084*, .

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*, 209–226.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bates, D., & Sarkar, D. (2007). The lme4 package. *R package version*, *2*.

Bebout, L. J., Segalowitz, S. J., & White, G. (1980). Children's comprehension of causal constructions with "because" and "so". *Child Development*, (pp. 565–568).

Benamara, F., & Taboada, M. (2015). Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics* (pp. 147–152). Denver, CL.

Besser, J., & Alexandersson, J. (2007). A comprehensive disfluency model for multiparty interaction. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue* (pp. 182–189). Antwerp, Belgium.

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Biran, O., & Rambow, O. (2011). Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, *5*, 363–381.

Black, J. B., & Bern, H. (1981). Causal coherence and memory for events in narratives. *Journal of Verbal Learning and Verbal Behavior*, *20*, 267–275.

Blakemore, D. (1997). Restatement and exemplification: A relevance theoretic reassessment of elaboration. *Pragmatics & Cognition*, *5*, 1–19.

Bloom, L., Lahey, M., Hood, L., Lifter, K., & Fiess, K. (1980). Complex sentences: Acquisition of syntactic connectives and the semantic relations they encode. *Journal of Child Language*, *7*, 235–261.

Blything, L. P., Davies, R., & Cain, K. (2015). Young children's comprehension of temporal relations in complex sentences: The influence of memory on performance. *Child Development*, *86*, 1922–1934.

van den Bosch, L. J., Segers, E., & Verhoeven, L. (2018). Online processing of causal relations in beginning first and second language readers. *Learning and Individual Differences*, *61*, 59–67.

Boudewyn, M. A. (2015). Individual differences in language processing: Electrophysiological approaches. *Language and Linguistics Compass*, *9*, 406–419.

Bourgonje, P., Grishina, Y., & Stede, M. (2017). Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it)* (pp. 1–6). Rome, Italy.

Brewer, W. F., & Lichtenstein, E. H. (1982). Stories are to entertain: A structural-affect theory of stories. *Journal of Pragmatics*, *6*, 473–486.

Britton, B. K. (1994). *Understanding expository text: Building mental structures to induce insights*. Academic Press.

van den Broek (2010). Using texts in science education: Cognitive processes and knowledge representation. *Science*, *328*, 453–456.

Bryant, J., Skolarus, L. E., Smith, B., Adelman, E. E., & Meurer, W. J. (2013). The accuracy of surrogate decision makers: Informed consent in hypothetical acute stroke scenarios. *BMC Emergency Medicine*, *13*, 18–24.

Bunt, H., & Prasad, R. (2016). Iso dr-core (iso 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)* (pp. 45–54). Portoroz, Slovenia.

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

Canestrelli, A. R., Mak, W. M., & Sanders, T. J. M. (2016). The influence of genre on the processing of objective and subjective causal relations: Evidence from eye-tracking. In *Genre in Language, Discourse and Cognition* (pp. 51–73). Walter de Gruyter.

Carlson, L., & Marcu, D. (2001). Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, *54*, 1–56.

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current and new directions in discourse and dialogue* (pp. 85–112). Springer.

Carston, R. (1993). Conjunction, explanation and relevance. *Lingua*, *90*, 27–48.

Chafe, W., & Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology*, *16*, 383–407.

Chiarcos, C. (2014). Towards interoperable discourse annotation: Discourse features in the Ontologies of Linguistic Annotation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)* (pp. 4569–4577). Reykjavik, Iceland.

Clark, E. V. (1971). On the acquisition of the meaning of *before* and *after*. *Journal of Verbal Learning and Verbal Behavior*, *10*, 266–275.

Clark, H. H., & Clark, E. V. (1968). Semantic distinctions and memory for complex sentences. *The Quarterly Journal of Experimental Psychology*, *20*, 129–138.

Clifton, C., Frazier, L., & Connine, C. (1984). Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, *23*, 696–708.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, R. (1987). Analyzing the structure of argumentative discourse. *Computational Linguistics*, *13*, 11–24.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*, 163–183.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*, 547–552.

Conway, A. R. A., Kane, M. J., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786.

Cornish, F. (2009). "Text'' and "discourse" as context. *Working Papers in Functional Discourse Grammar (WP-FDG-82): The London Papers I, 2009*, (pp. 97–115).

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

Crible, L., & Degand, L. (2017). Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory*, (pp. 1–26).

Crowle, C., Galea, C., Morgan, C., Novak, I., Walker, K., & Badawi, N. (2017). Inter-observer agreement of the general movements assessment with infants following surgery. *Early Human Development*, *104*, 17–21.

Cuenca, M.-J. (2003). Two ways to reformulate: A contrastive analysis of reformulation markers. *Journal of Pragmatics*, *35*, 1069–1093.

Cuenca, M.-J., & Marín, M.-J. (2009). Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics*, *41*, 899–914.

Dancygier, B., & Sweetser, E. (2000). Constructions with *if*, *since*, and *because*: Causality, epistemic stance, and clause order. *Topics in English Linguistics*, *33*, 111–142.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433.

Das, D., & Taboada, M. (2018). RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, *52*, 149–184.

De Kuthy, K., Ziai, R., & Meurers, D. (2016). Focus annotation of task-based data: Establishing the quality of crowd annotation. In *Proceedings of the Linguistic Annotation Workshop (LAW X)* (pp. 110–119). Berlin, Germany.

Degand, L. (1998). On classifying connectives and coherence relations. In *Proceedings of the 1998 ACL Workshop on Discourse Relations and Discourse Markers* (pp. 29–35). Montreal, Canada.

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

Demberg, V., Asr, F. T., & Scholman, M. C. J. (submitted). How consistent are our discourse annotations? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, (pp. 1–44).

Diessel, H. (2008). Iconicity of sequence: A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive Linguistics*, *19*, 465–490.

Diessel, H., & Hetterle, K. (2011). Causal clauses: A cross-linguistic investigation of their structure, meaning, and use. *Linguistic Universals and Language Variation*, (pp. 21–52).

Drenhaus, H., Demberg, V., Köhne, J., & Delogu, F. (2014). Incremental and predictive discourse markers: ERP studies on German and English. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci)* (pp. 403–408). Quebec, Canada.

Ehrlich, K. (1980). Comprehension of pronouns. *The Quarterly Journal of Experimental Psychology*, *32*, 247–255.

Elman, J. L., Kehler, A., & Rohde, H. (2006). Event structure and discourse coherence biases in pronoun interpretation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 697–702). Mahwah, NJ.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.

Evers-Vermeul, J., Hoek, J., & Scholman, M. C. J. (2017). On temporality in discourse annotation: Theoretical and practical considerations. *Dialogue & Discourse*, *8*, 1–20.

Evers-Vermeul, J., & Sanders, T. J. M. (2009). The emergence of Dutch connectives: How cumulative cognitive complexity explains the order of acquisition. *Journal of Child Language*, *36*, 829–854.

Evers-Vermeul, J., & Sanders, T. J. M. (2011). Discovering domains – on the acquisition of causal connectives. *Journal of Pragmatics*, *43*, 1645–1662.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*, 469–495.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*, *43*, 543–549.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS one*, *8*, 1–18.

Fisher, R. A. (1922). On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, *85*, 87–94.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*, 581–590.

Fuchs, S., Pape, D., Petrone, C., & Perrier, P. (2015). *Individual differences in speech production and perception* volume 3. Peter Lang Publishing Group.

Fuller, G., Kemp, S., & Raftery, M. (2017). The accuracy and reproducibility of video assessment in the pitch-side management of concussion in elite rugby. *Journal of Science and Medicine in Sport*, *20*, 246–249.

Gernsbacher, M. A. (1997). Coherence cues mapping during comprehension. *Processing interclausal relationships. Studies in the production and comprehension of text*, (pp. 3–22).

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*, 1–26.

González, M. (2005). Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies*, *7*, 53–86.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, *48*, 163–189.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395.

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

Gwet, K. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. STATAXIS Publishing Company.

Haberlandt, K., & Bingham, G. (1978). Verbs contribute to the coherence of brief narratives: Reading related and unrelated sentence triples. *Journal of Verbal Learning and Verbal Behavior*, *17*, 419–425.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Pittsburgh, PA.

Haley, D. (2009). *Applying latent semantic analysis to computer assisted assessment in the computer science domain: A framework, a tool, and an evaluation*. Ph.D. thesis The Open University.

Halliday, M. A. (1994). *Functional grammar*. London: Edward Arnold.

Halverson, S. (2004). Connectives as a translation problem. *An International Encyclopedia of Translation Studies*, *1*, 562–572.

Hannon, B., & Daneman, M. (1998). Facilitating knowledge-based inferences in less-skilled readers. *Contemporary Educational Psychology*, *23*, 149–172.

Hannon, B., & Daneman, M. (2001). Susceptibility to semantic illusions: An individual-differences perspective. *Memory & Cognition*, *29*, 449–461.

Hannon, B., & Daneman, M. (2004). Shallow semantic processing of text: An individual-differences account. *Discourse Processes*, *37*, 187–204.

Hayes-Roth, B., & Walker, C. (1979). Configural effects in human memory: The superiority of memory over external information sources as a basis for inference verification. *Cognitive Science*, *3*, 119–140.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, *56*, 473–493.

Hillard, D., Purpura, S., & Wilkerson, J. (2007). An active learning framework for classifying political text. In *Annual Meeting of the Midwest Political Science Association*. Chicago, Ill.

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

Hobbs, J. R. (2003). Discourse and inference. Unpublished manuscript.

Hoek, J. (2018). *Making sense of discourse: On discourse segmentation and the linguistic marking of coherence relations*. Ph.D. thesis Utrecht University.

Hoek, J., Evers-Vermeul, J., & Sanders, T. J. M. (2017a). Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, *14*, 357–386.

Hoek, J., & Scholman, M. C. J. (2017). Evaluating discourse annotation: Some recent insights and new approaches. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)* (pp. 1–13). Toulouse, France.

Hoek, J., & 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)* (pp. 39–45). London, UK.

Hoek, J., Zufferey, S., Evers-Vermeul, J., & Sanders, T. J. M. (2017b). Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, *121*, 113–131.

Horowitz, R., & Samuels, S. J. (1987). *Comprehending Oral and Written Language*. ERIC.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., & Hovy, E. (2013). Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 1120–1130). Denver, CO.

Hovy, E. H. (1995). The multifunctionality of discourse markers. In *Proceedings of the Workshop on Discourse Markers* (pp. 1–11). Egmond aan Zee, the Netherlands.

Hovy, E. H., & Maier, E. (1995). Parsimonious or profligate: How many and which discourse structure relations? Unpublished manuscript.

Hyland, K. (2007). Applying a gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics*, *28*, 266–285.

Izutsu, M. N. (2008). Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations. *Journal of Pragmatics*, *40*, 646–675.

Jansen, P., Surdeanu, M., & Clark, P. (2014). Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 977–986). Baltimore, MD.

Jasinskaja, K., & Karagjosova, E. (2011). Elaboration and explanation. *Constraints in Discourse*, *4*, 1–11.

Johnson, N. S. (1983). What do you do if you can't tell the whole story? the development of summarization skills. *Children's Language*, *4*, 315–383.

Just, M. A., & Carpenter, P. A. (1978). Inference processes during reading: Reflections from eye fixations. In J. Senders, D. Fisher, & R. Monty (Eds.), *Eye movements and the higher psychological functions* (pp. 157–174). Erlbaum Hillsdale, NJ.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Kaiser, E. (2012). Taking action: A cross-modal investigation of discourse-level representations. *Frontiers in Psychology*, *3*, 156.

Kallen, J. L., & Kirk, J. M. (2012). *SPICE-Ireland: A User's Guide*. Belfast: Cló Ollscoil na Banríona.

Kamalski, J., Sanders, T. J. M., & Lentz, L. (2008). Coherence marking, prior knowledge, and comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes*, *45*, 323–345.

Kawahara, D., Machida, Y., Shibata, T., Kurohashi, S., Kobayashi, H., & Sassano, M. (2014). Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of the International Conference on Computational Linguistics (COLING)* (pp. 269–278). Dublin, Ireland.

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

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

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

Kennedy, W. (2009). Cognitive plausibility in cognitive modeling. In *Artificial Intelligence, and Social Simulation. 9th International Conference on Cognitive Modeling* (pp. 454–455). Manchester, UK.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2017). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, (pp. 1–16).

Knoepke, J., Richter, T., Isberner, M.-B., Naumann, J., Neeb, Y., & Weinert, S. (2017). Processing of positive-causal and negative-causal coherence relations in primary school children and adults: A test of the cumulative cognitive complexity approach in German. *Journal of Child Language*, *44*, 297–328.

Knott, A. (1996). *Motivating a Set of Coherence Relations*. Ph.D. thesis University of Edinburgh.

Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, *18*, 35–62.

Knott, A., & Sanders, T. J. M. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, *30*, 135–175.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press.

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

Koornneef, A. W., & 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*, 445–465.

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., & Staudacher, M. (2006). Measuring and reconstructing pointing in visual contexts. In *Proceedings of the Brandial* (pp. 82–89). Potsdam, Germany.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*, 411–433.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? what the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, *31*, 602–616.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*, 32–59.

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

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389–433.

Lascarides, A., Asher, N., & Oberlander, J. (1992). Inferring discourse relations in context. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics* (pp. 1–8). Newark, DE.

Lehman, S., & Schraw, G. (2002). Effects of coherence and relevance on shallow and deep text processing. *Journal of Educational Psychology*, *94*, 738–750.

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

Li, J. J., & Nenkova, A. (2016). The Instantiation discourse relation: A corpus analysis of its properties and improved detection. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 1181–1186). San Diego, CA.

Linderholm, T., Everson, M. G., Van Den Broek, P., Mischinski, M., Crittenden, A., & Samuels, J. (2000). Effects of causal text revisions on more- and less-skilled readers' comprehension of easy and difficult texts. *Cognition and Instruction*, *18*, 525–556.

Louis, A., Joshi, A., & Nenkova, A. (2010). Discourse indicators for content selection in summarization. In *The 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 147–156). Tokyo, Japan.

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Magliano, J. P., Baggett, W. B., Johnson, B. K., & Graesser, A. C. (1993). The time course of generating causal antecedent and causal consequence inferences. *Discourse Processes*, *16*, 35–53.

Magliano, J. P., & Graesser, A. C. (1991). A three-pronged method for studying inference generation in literary text. *Poetics*, *20*, 193–232.

Mak, W. M., & Sanders, T. J. M. (2013). The role of causality in discourse processing: Effects of expectation and coherence relations. *Language and Cognitive Processes*, *28*, 1414–1437.

Mak, W. M., Tribushinina, E., & Andreiushina, E. (2013). Semantics of connectives guides referential expectations in discourse: An eye-tracking study of Dutch and Russian. *Discourse Processes*, *50*, 557–576.

von der Malsburg, T., Poppels, T., & Levy, R. P. (2018). Implicit gender bias in linguistic descriptions for expected events. *PsyArXiv*, .

von der Malsburg, T., & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*, 1545–1578.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*, 243–281.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Marks, D., Comans, T., Thomas, M., Ng, S. K., O'Leary, S., Conaghan, P. G., Scuffham, P. A., & Bisset, L. (2016). Agreement between a physiotherapist and an orthopaedic surgeon regarding management and prescription of corticosteroid injection for patients with shoulder pain. *Manual Therapy*, *26*, 216–222.

Martins, D., Kigiel, D., & Jhean-Larose, S. (2006). Influence of expertise, coherence, and causal connectives on comprehension and recall of an expository text. *Current Psychology Letters. Behaviour, Brain & Cognition*, *3*.

Maschler, Y., & Schiffrin, D. (2015). Discourse markers: Language, meaning, and context. *The Handbook of Discourse Analysis*, *2*, 189–221.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1–43.

Meyer, T., & Popescu-Belis, A. (2012). Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)* (pp. 129–138). Avignon, France.

Miller, G. A. (1969). A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology*, *6*, 169–191.

Moore, J. D., & Pollack, M. E. (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, *18*, 537–544.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., & Tily, H. (2010). Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 122–130). Los Angeles, CA.

Münte, T. F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, *395*, 71–73.

Murray, J. D. (1997). Connectives and narrative text: The role of continuity. *Memory & Cognition*, *25*, 227–236.

Myers, J. L., Shinjo, M., & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language*, *26*, 453–465.

Noordman, L., & Rijswijk, W. v. (1997). De functie van het voegwoord "hoewel" voor de samenhang van tekst. *Tijdschrift voor Taalbeheersing*, *3*, 252–264.

Noordman, L. G., & Vonk, W. (2015). Inferences in discourse, psychology of. In *International Encyclopedia of the Social & Behavioral Sciences (2nd ed.) Vol. 12* (pp. 37–44). Elsevier.

Noordman, L. G., Vonk, W., & Kempff, H. J. (1992). Causal inferences during the reading of expository texts. *Journal of Memory and Language*, *31*, 573–590.

Noordman, L. G. M., & de Blijzer, F. (2000). On the processing of causal relations. In *Cause - Condition - Concession - Contrast: Cognitive and Discourse Perspectives* (pp. 35–56). Walter de Gruyter.

Noordman, L. G. M., & Vonk, W. (1998). Memory-based processing in understanding causal information. *Discourse Processes*, *26*, 191–212.

Nuyts, J. (1992). *Aspects of a Cognitive-pragmatic Theory of Language: On Cognition, Functionalism and Grammar*. John Benjamins Publishing Company.

Oakhill, J. (1984). Inferential and memory skills in children's comprehension of stories. *British Journal of Educational Psychology*, *54*, 31–39.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*. Springer Verlag.

Pander Maat, H. (1998). Classifying negative coherence relations on the basis of linguistic evidence. *Journal of Pragmatics*, *30*, 177–204.

Pander Maat, H., & Sanders, T. J. M. (2000). Domains of use or subjectivity? the distribution of three Dutch causal connectives explained. *Topics in English Linguistics*, *33*, 57–82.

Park, J., & 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* (pp. 108–112). Seoul, South Korea.

Passonneau, R. J., & Carpenter, B. (2014). The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, *2*, 311–326.

Peldszus, A., & Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *7*, 1–31.

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*, 1824–1854.

Pitler, E., Louis, A., & 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* (pp. 683–691). Suntec, Singapore.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., & Joshi, A. K. (2008). Easily identifiable discourse relations. Technical report.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, *12*, 601–638.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Marocco.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A. K., Robaldo, L., & Webber, B. (2007). *The Penn Discourse Treebank 2.0 annotation manual*.

Prasad, R., Riley, K. F., & Lee, A. (2017). Towards full text shallow discourse relation annotation: Experiments with cross-paragraph implicit relations in the PDTB. *Proceedings of the SIGdial 2017 Conference*, (pp. 7–16).

Prasad, R., Webber, B., & Joshi, A. (2014). Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*, *40*, 921–950.

Purpura, S., & Hillard, D. (2006). Automated classification of congressional legislation. In *Proceedings of the 2006 International Conference on Digital Government Research* (pp. 219–225). San Diego, CA.

Pusse, F., Sayeed, A., & Demberg, V. (2016). LingoTurk: Managing crowdsourced tasks for psycholinguistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 57–61). San Diego, CA.

Pyykkönen, P., & Järvikivi, J. (2012). Children and situation models of multiple events. *Developmental Psychology*, *48*, 521.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `http://www.R-project.org` ISBN 3-900051-07-0.

Rapp, D. N., Broek, P. v. d., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading*, *11*, 289–312.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.

Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, *14*, 367–381.

Rehbein, I., Scholman, M. C. J., & Demberg, V. (2016). Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)* (pp. 23–28). Portoroz, Slovenia.

Riezler, S. (2014). On the problem of theoretical terms in empirical computational linguistics. *Computational Linguistics*, *40*, 235–245.

Robaldo, L., & Miltsakaki, E. (2014). Corpus-driven semantics of concession: Where do expectations come from? *Dialogue & Discourse*, *5*, 1–36.

Rohde, H., Dickinson, A., Clark, C., Louis, A., & Webber, B. (2015). Recovering discourse relations: Varying influence of discourse adverbials. In *Proceedings of the Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)* (pp. 1–22). Lisbon, Portugal.

Rohde, H., Dickinson, A., Schneider, N., Clark, C., Louis, A., & Webber, B. (2016). Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW X)* (pp. 49–58). Berlin, Germany.

Rohde, H., & Horton, W. S. (2014). Anticipatory looks reveal expectations about discourse relations. *Cognition*, *133*, 667–691.

Rutherford, A., & Xue, N. (2014). Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 645–654). Gothenburg, Sweden.

Rutherford, A., & Xue, N. (2015). Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 799–808). Denver, CO.

Sanders, T. J. M. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes*, *24*, 119–147.

Sanders, T. J. M. (2005). Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning* (pp. 105–114). Toulouse, France.

Sanders, T. J. M. (2017). Do we seek for causality in discourse? On the cognition of coherence relations and connectives. In *Proceedings of 50th Annual Meeting of the Societas Linguistica Europaea (SLE)* (p. 397). Zürich, Switzerland.

Sanders, T. J. M., Demberg, V., Hoek, J., Scholman, M. C. J., Torabi Asr, F., Zufferey, S., & Evers-Vermeul, J. (2018). Unifying dimensions in discourse relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, *ahead of print*, 1–71.

Sanders, T. J. M., & Evers-Vermeul, J. (to appear). Subjectivity and causality in discourse and cognition; Evidence from corpus analyses, acquisition and processing. In O. Loureda, I. R. Fernández, L. Nadal, & A. Cruz (Eds.), *Methodological approaches to discourse markers* (pp. 1–26). John Benjamins.

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, *29*, 37–60.

Sanders, T. J. M., & Spooren, W. (1999). Communicative intentions and coherence relations. In W. Bublits, U. Lenk, & E. Ventola (Eds.), *Coherence in Spoken and Written Discourse* (pp. 235–250). John Benjamins.

Sanders, T. J. M., Spooren, W. P., & Noordman, L. G. (1993). Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics*, *4*, 93–133.

Sanders, T. J. M., Spooren, W. P. M. S., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, *15*, 1–35.

Sanders, T. J. M., Vis, K., & Broeder, D. (2012). Project notes of CLARIN project DiscAn: Towards a discourse annotation system for Dutch language corpora. In *Proceedings 8th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA)* (pp. 61–65). Pisa, Italy.

Sanford, A. J., & Graesser, A. C. (2006). Shallow processing and underspecification. *Discourse Processes*, *42*, 99–108.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*, 382–386.

Scholman, M. C. J., & Demberg, V. (2017a). Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop (LAW)* (pp. 24–33). Valencia, Spain.

Scholman, M. C. J., & Demberg, V. (2017b). Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue & Discourse*, *8*, 56–83.

Scholman, M. C. J., Evers-Vermeul, J., & Sanders, T. J. M. (2016). Categories of coherence relations in discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse*, *7*, 1–28.

Scholman, M. C. J., Rohde, H., & Demberg, V. (2017). "On the one hand" as a cue to anticipate upcoming discourse structure. *Journal of Memory and Language*, *97*, 47–60.

Scott, W. A. (1959). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321–325.

Segal, E. M., Duchan, J. F., & Scott, P. J. (1991). The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse Processes*, *14*, 27–54.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Sharp, R., Jansen, P., Surdeanu, M., & Clark, P. (2015). Spinning straw into gold: Using free text to train monolingual alignment models for non-factoid question answering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 231–237). Denver, CO.

Shi, W., & Demberg, V. (2017). Do we need cross validation for discourse relation classification? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 150–156). Valencia, Spain.

van Silfhout, G., Evers-Vermeul, J., Mak, W. M., & Sanders, T. J. (2014). Connectives and layout as processing signals: How textual features affect students' processing and text representation. *Journal of Educational Psychology*, *106*, 1036–1048.

van Silfhout, G., Evers-Vermeul, J., & Sanders, T. J. M. (2015). Connectives as processing signals: How students benefit in processing narrative and expository texts. *Discourse Processes*, *52*, 47–76.

Simner, J., & Pickering, M. J. (2005). Planning causes and consequences in discourse. *Journal of Memory and Language*, *52*, 226–239.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 254–263). Waikiki, HI.

Song, L. (2010). The role of context in discourse analysis. *Journal of Language Teaching and Research*, *1*, 876–879.

Spooren, W. (1997). The processing of underspecified coherence relations. *Discourse Processes*, *24*, 149–168.

Spooren, W. P. M. S., & Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, *6*, 241–266.

Stab, C., & Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 46–56). Doha, Qatar.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, (pp. 402–433).

Staub, A., & Clifton, C. (2006). Syntactic prediction in language comprehension: Evidence from *either...or*. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 425–436.

Steele, D., & Specia, L. (2014). Divergences in the usage of discourse markers in English and Mandarin Chinese. In *International Conference on Text, Speech, and Dialogue* (pp. 189–200). Brno, Czech Republic.

Stein, B. L., & Kirby, J. R. (1992). The effects of text absent and text present conditions on summarization and recall of text. *Journal of Reading Behavior*, *24*, 217–232.

Stevenson, R., Knott, A., Oberlander, J., & McDonald, S. (2000). Interpreting pronouns and connectives: Interactions among focusing, thematic roles and coherence relations. *Language and Cognitive Processes*, *15*, 225–262.

Stukker, N., & Sanders, T. J. M. (2012). Subjectivity and prototype structure in causal connectives: A cross-linguistic perspective. *Journal of Pragmatics*, *44*, 169–190.

Sun, K., & Zhang, L. (2018). Quantitative aspects of PDTB-style discourse relations across languages. *Journal of Quantitative Linguistics*, *25*, 342–371.

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability – and a little bit more. *Intelligence*, *30*, 261–288.

Taboada, M., & Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, *4*, 249–281.

Taboada, M., & Mann, W. C. (2006). Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, *8*, 423–459.

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, (pp. 1–21).

Tonelli, S., Riccardi, G., Prasad, R., & Joshi, A. K. (2010). Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)* (pp. 2084–2090). Marrakesh, Marocco.

Townsend, D. J. (1983). Thematic processing in sentences and texts. *Cognition*, *13*, 223–261.

Trabasso, T., & Van Den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language*, *24*, 612–630.

Traxler, M. J., Sanford, A. J., Aked, J. P., & Moxey, L. M. (1997). Processing causal and diagnostic statements in discourse. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 88.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 443–467.

Van Veen, R. (2011). *The acquisition of causal connectives: The role of parental input and cognitive complexity*. Ph.D. thesis Utrecht University.

van Veen, R., Evers-Vermuel, J., Sanders, T. J. M., & van den Bergh, H. (2014). Why? because i?m talking to you! parental input and cognitive complexity as determinants of children?s connective acquisition. In *The pragmatics of discourse coherence: Theories and applications*. Amsterdam: John Benjamins.

Vergez-Couret, M., & Adam, C. (2012). Signaling Elaboration: Combining french gerund clauses with lexical cohesion cues. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, *10*, 1–22.

Versley, Y. (2011). Multilabel tagging of discourse relations in ambiguous temporal connectives. In *International Conference on Recent Advances in Natural Language Processing (RANLP)* (pp. 154–161). Hissar, Bulgaria.

Versley, Y., & Gastel, A. (2013). Linguistic tests for discourse relations in the TüBa-D/Z corpus of written German. *Dialogue & Discourse*, *4*, 142–173.

Waters, G., Caplan, D., & Hildebrandt, N. (1987). Working memory and written sentence comprehension. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 531–555). Lawrence Erlbaum Associates, Inc.

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology Section A*, *49*, 51–79.

Webber, B. (2013). What excludes an alternative in coherence relations. In *Proceedings of the International Conference on Computational Semantics (IWCS)* (pp. 921–950). Potsdam, Germany.

Webber, B., Knott, A., & Joshi, A. (2001). Multiple discourse connectives in a lexicalized grammar for discourse. In *Computing Meaning* (pp. 229–245). Springer.

Webber, B., Knott, A., Stone, M., & 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* (pp. 41–48). College Park, MD.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW X)* (pp. 22–31). Berlin, Germany.

Wei, Y. (2018). *Causal connectives and perspective markers in Chinese: The encoding and processing of subjectivity in discourse*. Ph.D. thesis Utrecht University.

Wolf, F., & Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, *31*, 249–287.

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*, 648–672.

Xu, X., Jiang, X., & 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*, 704–715.

Xue, N. (2005). Annotating discourse connectives in the Chinese Treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky* (pp. 84–91). Ann Arbor, MI.

Xue, N., Ng, H. T., Pradhan, S., Prasad, R., Bryant, C., & Rutherford, A. (2015). The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task* (pp. 1–16). Beijing, China.

Xue, N., Ng, H. T., Pradhan, S., Rutherford, A., Webber, B., Wang, C., & Wang, H. (2016). Conll 2016 shared task on multilingual shallow discourse parsing. *Proceedings of the CoNLL-16 shared task*, (pp. 1–19).

Yang, H., Callan, J., & Shulman, S. (2006). Next steps in near-duplicate detection for eRulemaking. In *Proceedings of the 2006 International Conference on Digital Government Research* (pp. 239–248). San Diego, CA.

Ye, Z., Kutas, M., George, M. S., Sereno, M. I., Ling, F., & Münte, T. F. (2012). Rearranging the world: Neural network supporting the processing of temporal connectives. *NeuroImage*, *59*, 3662–3667.

Yoshida, M., Dickey, M. W., & Sturt, P. (2013). Predictive processing of syntactic structure: Sluicing and ellipsis in real-time sentence processing. *Language and Cognitive Processes*, *28*, 272–302.

Zhou, Y., & Xue, N. (2015). The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, *49*, 397–431.

Zufferey, S., & Cartoni, B. (2014). A multifactorial analysis of explicitation in translation. *Target. International Journal of Translation Studies*, *26*, 361–384.

Zufferey, S., & Degand, L. (2013). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, *1*, 1–24.

Zufferey, S., Degand, L., Popescu-Belis, A., & Sanders, T. J. M. (2012). Empirical validations of multilingual annotation schemes for discourse relations. In *Proceedings of the 8th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-8)* (pp. 77–84). Jeju Island, Korea.

Zufferey, S., Mak, W. M., & Sanders, T. J. M. (2015). A cross-linguistic perspective on the acquisition of causal connectives and relations. *International Review of Pragmatics*, *7*, 22–39.