The Theory of Correlation Formulas and Their Application to Discourse Coherence

Julian Michael

Thesis submitted to The University of Texas at Austin
in partial fulfillment of the requirements for graduation with
the Dean's Scholars Honors Degree in Computer Science

Vladimir Lifschitz Supervising Professor	Date
Robert van de Geijn Honors Advisor in Computer Science	Date

Acknowledgments

First and foremost, I'd like to thank Vladimir Lifschitz, who supervised me in my undergraduate research and for this thesis. In the two years I have been in his group, I have learned immeasurably. From our discussions, I took away lessons about the value of precision, the meaning of logical proofs, the nature of clear thought, and how to conduct research. He also has shown me how to create a productive research environment—the combination of a patient and welcoming atmosphere with careful rigor and attention to details is something I hope I can even nearly replicate in my future research career. His patience and input was also invaluable in the (admittedly hurried) writing of this thesis—many of the positive qualities of this work are due to his insights, while any faults or fallacies are my own.

I also thank the co-authors of our joint work on the Correlation Calculus—Daniel Bailey, Amelia Harrison, and Yuliya Lierler—for making this thesis possible. I also greatly appreciate Amelia's comments on a draft of this thesis, and Vladimir and Yuliya have both given me valuable advice about my research career.

Katrin Erk has also served as a valuable mentor, always introducing me to new ideas in semantics and helping me figure out my direction in research. I also would like to thank Hans Kamp, who has taught me a great deal about formal semantics and Discourse Representation Theory.

Finally, I extend thanks to my friends and family, who have supported me when I needed it through my college career, and have made me who I am today.

Abstract

The Winograd Schema Challenge (WSC) was proposed as a measure of machine intelligence. It boils down to *anaphora resolution*, a task familiar from computational linguistics. Research in linguistics and AI has coalesced around *discourse coherence* as the critical factor in solving this task, and the process of establishing discourse coherence relies fundamentally on world and commonsense knowledge.

In this thesis, we build on an approach to establishing coherence on the basis of *correlation*. The utility of this approach lies in its conceptual clarity and ability to flexibly represent commonsense knowledge. We work to fill some conceptual holes with the Correlation Calculus approach. First, understanding the calculus in a vacuum is not straightfoward unless it has a precise semantics. Second, existing demonstrations of the Correlation Calculus on Winograd Schema Challenge problems have not been linguistically credible.

We hope to ameliorate some—but by no means all—of the outstanding issues with the Correlation Calculus. We do so first by providing a precise semantics of the calculus, which relates our intuitive understanding of correlation with a precise notion involving probabilities. Second, we formulate the establishment of discourse coherence by correlation formulas within the framework of Discourse Representation Theory. This provides a more complete and linguistically credible account of the relationship between the Correlation Calculus, discourse coherence, and Winograd Schema Challenge problems.

Contents

Chapter 1

Introduction

The Winograd Schema Challenge was proposed by Levesque *et al.* [2012] as a measure of machine intelligence. It boils down to *anaphora resolution*, a task familiar from computational linguistics. Research in linguistics and AI has coalesced around *discourse coherence* as the critical factor in solving this task, and the process of establishing discourse coherence relies fundamentally on world and commonsense knowledge.

Bailey *et al.* [2015] proposed an approach to establishing coherence on the basis of *correlation*. The utility of this approach lies in its conceptual clarity and ability to flexibly represent commonsense knowledge. They exhibit a use of the calculus in an application to the Winograd Schema Challenge. However, there are some shortcomings to their approach. Understanding the calculus in a vacuum is not straightfoward—for example, the calculus on its own does not give a test of non-derivability; we cannot verify that our intuition of the calculus is accurate unless it has a precise semantics.[1] Another problem with the approach presented in Bailey *et al.* [2015] was that the formulation of Winograd Schema Challenge problems was not linguistically credible. It involves a translation from natural language clauses in WSC problems to formulas of First-Order Logic, but the translation was not based on rigorous principles, so there was not evidence that a necessary component of the reasoning behind the WSC problem was not happening in that step.

In this thesis, we hope to ameliorate some—but by no means all—of the outstanding issues with the Correlation Calculus. We do so first by providing a precise

[1] A semantics for the calculus was given in Bailey *et al.* [2015], but since it was primarily a contribution of the author of this thesis, it is included here as a contribution.

semantics of the calculus, which relates our intuitive understanding of correlation with a precise notion involving probabilities. Second, we formulate the establishment of discourse coherence by correlation formulas within the framework of Discourse Representation Theory. This provides a more complete and linguistically credible account of the relationship between the Correlation Calculus, discourse coherence, and Winograd Schema Challenge problems.

There are still many outstanding issues when it comes to solving the Winograd Schema Challenge. A paramount problem—which we do not address—is how to obtain the knowledge required for deriving or explaining solutions to Winograd Schema Challenge problems. This is an important direction for future work, but clarity in the *uses* of the Correlation Calculus in discourse coherence does go towards clarifying how to represent commonsense knowledge about correlation.

This thesis is organized as follows. We will first cover necessary background information in Chapter 2, with discussions of the Winograd Schema Challenge, the Correlation Calculus, and Discourse Representation Theory. Then, in Chapter 3, we will discuss a probabilistic semantics for the Correlation Calculus. In Chapter 4 we develop how to relate the calculus with Discourse Representation Theory in the process of establishing coherence, and we conclude with a discussion of possible problems and future directions in Chapter 5.

Chapter 2

Background

2.1 The Winograd Schema Challenge

The Winograd Schema Challenge (WSC) was proposed by Levesque *et al.* [2012] as a measure of machine intelligence. It was designed to mitigate several cited problems with the original Turing Test:

- Deception—the machine has to create a false identity in order to pose as a human.
- Conversation—a lot of interactions may qualify as "legitimate conversation" and yet not draw on intelligent reasoning.
- Evaluation—different judges may come to different verdicts, and often the real result is inconclusive.

A better alternative, Levesque *et al.* argued, would be a test of comprehension of a broad range of natural language sentences that could be administered and scored by a machine.

The test they proposed boils down to *anaphora resolution*, a task familiar from computational linguistics where the goal is to determine the antecedent of some anaphoric element of a sentence—in this case, a pronoun. *Winograd Schemas* are sentence schemas with two instances. They are named after Terry Winograd for his example from Winograd [1972]:

(1) a. The city councilmen denied the protestors a permit because they advocated violence.

b. The city councilmen denied the protestors a permit because they feared violence.

A Winograd Schema Challenge question continues:

Who [advocated/feared] violence?

In the case of (1a), the answer is *the protestors*, and in the case of (1b), it is *the city councilmen*; either way the question is of the antecedent of *they*. It was noted by Winograd [1972] that resolving this pronoun requires knowledge and reasoning ability. The general *structure* of the sentences has a lot to do with this.

First, each sentence contains two independent clauses, neither of which is subordinate to the other. This means that grammatical binding restrictions—based on clause-internal patterns such as subjectivity or nuclearity—may not constrain the antecedent of the pronoun. Second, the grammatical structure of the sentence is the same in either case, as just one word is switched out. This means that syntactic patterns—which are used by state-of-the-art systems for coreference (Lee *et al.* [2011], Lee *et al.* [2013])—are of no use in this problem, and statistics on words and phrases may not be helpful in any obvious way. Together, these controls result in the fact that in many sentences of this general form, it seems that commonsense reasoning or world knowledge is required to resolve the anaphor. This is borne out in examples such as Winograd's: it seems that the reason for the resolvent of the anaphor has to do with general knowledge about protestors, councilmen, and the reasons for public demonstration.

So the Winograd Schema Challenge consists of a series of anaphora resolution problems that have the same general form as Winograd's example. Formally, a Winograd Schema consists of

- A sentence or brief discourse in English.[1] It must contain:
  - two noun phrases of the same general semantic class (male, female, inanimate, or group of people/objects),
  - a pronoun (possibly in possessive form) referring to one of the noun phrases, and
  - a "special word" and "alternate word" such that if the special word is replaced with the alternate word, the referent of the pronoun changes.

[1] We use the term "discourse" simply to refer to a sequence of sentences.

- A question echoing the condition on the pronoun in question.
- A pair of answers corresponding to the two noun phrases in the discourse, in their order of appearance.

Here are two complete examples from Levesque *et al.* [2012]:

- The trophy doesn't fit into the brown suitcase because it's too [big/small]. What was too [big/small]?

  Answer 0: The trophy

  Answer 1: The suitcase

- Joan made sure to thank Susan for all the help she had [received/given]. Who had [received/given] the help?

  Answer 0: Joan

  Answer 1: Susan

The schemas on the actual test may not correspond quite exactly to the definition, in the sense that instead of a special and alternate word, there may be a short special and alternate *phrase*, which, when switched out for each other, do not change the overall structure of the sentence.

The administration of the test would be as follows (as presented by Levesque [2014]):

1. Select N Winograd Schema questions,
2. randomly use the special or alternate word in the question,
3. present the test to the subject, obtaining N replies, and
4. assign a final grade of
$$\frac{\max(0, N - k \cdot \mathit{Wrong})}{N},$$

where *Wrong* is the number of incorrect answers and k is a penalty coefficient (to counteract guessing). The problems would be presented in a standard format that is explicit about the answer choices, so each problem is simply a 2-option multiple choice question. The bar for an intelligent AI would be performance comparable to a human baseline.

2.1.1 The Status of the WSC

Since its proposal, the WSC has seen some attention from the research community. Nuance Communications, Inc. is sponsoring a competition in Fall 2015, with a grand prize of $\$25,000$.

The test was designed with the idea of being "Google-proof," in the sense that the problems could not be solved using statistical information about the cooccurrence of words. Ernest Davis, in his compilation of Winograd Schemas,[2] occasionally justifies this with examples by giving the number of results returned when the relevant phrases are Google-searched.

This idea of "Googling" to find the answer is very much a simplification of the statistical machine learning techniques already being used in the NLP community. Many NLP researchers seem to believe that the WSC will not be a problem for machine learning methods. For example, Rahman and Ng [2012] apply exactly the sort of methods the test was designed to guard against and achieve accuracy of about 73% on their set of twin sentences designed after Winograd Schemas. But taking a look at their data reveals that it at times differs from Winograd Schemas in significant ways. Many can be resolved without an understanding of the proposition denoted by the sentence, for example:

> NATO forces did precision predator strikes for the Libyan rebels since they [had the predator vehicles/did not have the capabilities themselves].

This can be explained simply by very simple statistical data—NATO is likely to have predator vehicles, while Libyan rebels are more likely to be incapable. The entire proposition needs to be relevant to the solution in order to ensure that knowledge and reasoning play their part. This problem is listed as "Pitfall 1" by Levesque *et al.* [2012]—an answer too obvious. Others can even be solved by grammatical binding constraints—for example, the following schema instance:

> The mermaid swam toward Sue and made her gasp.

If "her" meant the mermaid it would have to have been "herself." We will not be undertaking a careful analysis of their data set, but these examples were found in just a very small sample. So their results may not be applicable to the Winograd Schema Challenge.

[2] See `https://www.cs.nyu.edu/davise/papers/WS.html`.

As it stands, the jury still seems to be out on whether the WSC can be solved using statistical or machine learning models that do not incorporate world knowledge in any obvious way. However, this is a different question than that of whether solving the WSC *requires* knowledge. The feeling in the logic-based KRR community seems to be that a problem that requires *world knowledge* requires *axioms* in the methodology of its solution. If we are to buy into Levesque's arguments and trust that the WSC *is indeed* a test of commonsense reasoning and world knowledge, then we must be prepared to admit that if a statistical system solves the Winograd Schema Challenge, then in a very clear sense, it has world knowledge and is performing commonsense reasoning.

But the value of the WSC is not just in its *evaluative* power. We have plenty of tasks in AI that encapsulate fragments of human reasoning abilities, which serve as powerful evaluations for our AI systems. And yet when a new system breaks yet another record, what have we learned about intelligence in the process? Levesque [2014] argues that the "test-driven" approach to AI has a habit of producing *idiot savants*. When we are focused just on passing a test, we may (indeed, we are incentivized to) miss the forest for the trees and develop a skill that does not generalize. So we should not view the WSC as a *test to be passed*. Instead, we should view human behavior on the WSC as a *phenomenon to be explained*.

So how do we explain human behavior on the challenge? Levesque [2014] presents a compelling argument that knowledge is crucial in how we come to the conclusions we do on Winograd Schema questions. It takes the form of a thought experiment, which we will abridge and repeat here. Consider the following hypothetical Winograd Schema instance:

(2) The large ball crashed through the table because it was made of XYZZY. What was made of XYZZY?

- the large ball
- the table

Here the answer is unclear. But what if we were told the following:

> The material XYZZY is ninety-eight percent air, making it lightweight and buoyant.

Now it's reasonably clear that *the table* is the answer to this question. In fact, this is a modification of a Winograd Schema with special and alternate words "steel"

and "styrofoam." The point here is that the above fact about styrofoam really may be behind the justification of the answer to the Winograd Schema question.

So several critical questions arise about the knowledge we need for the WSC, including:

1. How do we represent the knowledge?
2. How do we obtain the knowledge?
3. How do we use the knowledge?

Any complete account of how to intelligently solve or explain Winograd Schemas would have to address all of these issues. Our goal is not focused so much on developing an end-to-end system for solving the Winograd Schema Challenge. Instead, we aim to *explain* a small part of the *phenomena* of Winograd Schemas in a principled way. In particular, we focus on issue 3: how to use relevant commonsense knowledge to justify solutions to Winograd Schema questions. This will also have implications for issue 1. But our main concern in this thesis will be the interaction of commonsense knowledge and the *coherence* of a discourse.

2.2 Discourse Coherence

While our intuitions indicate that world knowledge is at play in the Winograd Schema Challenge, as Levesque [2014] aptly illustrated (see Section 2.1.1), defining *how* it factors into the decision in general is not straightforward. However, research in linguistics and AI has coalesced around *discourse coherence* as the critical factor.

Coherence may be defined differently by different groups, but it always corresponds roughly to the discourse "making sense." More specifically, it has to do with how the informational content of sentences in a discourse are bound together. Hobbs [1979] suggested that coreference resolution was a byproduct of the process of establishing discourse coherence, in the form of "petty implicatures"—assumptions of identity of two entities—that would be made in order to support the existence of "coherence relations" between sentences. Asher and Lascarides [2003] use a similar approach in Segmented Discourse Representation Theory, where these bonds are called *rhetorical relations* or *discourse relations* (though they are truly the same beasts, some even with the same names as Hobbs's relations). Alternatively, in the *interpretation as abduction* approach of Hobbs *et al.* [1993] coherence is enforced

by the "best explanation," which is a proof of the discourse content (from existing knowledge) that minimizes assumptions and redundancies.

Coherence has been implicated in a wide variety of disambiguation tasks, like lexical (i.e., word sense) disambiguation (Asher and Lascarides [1995]) and temporal anaphora resolution (Lascarides and Asher [1993]). But most notable is its role in coreference (and anaphora) resolution. Kehler *et al.* [2008] found that discourse coherence not only is the determining factor in coreference resolution, but it also explains syntax-based biases previously suggested in the literature having to do with subjecthood, grammatical roles, and thematic roles.

2.2.1 Determining Coherence

We will discuss a couple of established methods of determining coherence before moving on to the approach that inspired this work.

Interpretation as abduction First is the "interpretation as abduction" approach of Hobbs *et al.* [1993]. Under this approach, the entire process of discourse interpretation (which, roughly speaking, may be thought of as general language understanding) is modeled as an *abductive reasoning* problem. Abductive reasoning, or *inference to the best explanation*, means that the interpreter constructs a proof of the proposed content, using existing knowledge or making assumptions where necessary. The "best" explanation will generally make few assumptions and minimize redundancies where possible.

The theory has been implemented in *weighted abduction* systems, where the costs of abductive proofs are the sum of the costs of all of their assumptions. Under this framework, Inoue *et al.* [2012] have cast the abductive reasoning problem as an Integer Linear Programming problem, facilitating the use of weighted abduction—and knowledge—for coreference resolution. They report that they can improve the results of the Stanford resolver in certain situations. Whether their approach could be applied successfully to the WSC is not clear, but the method does provide a way to bring knowledge into the picture.

Segmented DRT Another approach to establishing coherence is in the framework of Segmented Discourse Representation Theory (SDRT), outlined by Asher and Lascarides [2003]. SDRT characterizes coherence by a set of rhetorical relations, and disambiguation prefers the interpretation that supports the greatest *quality*

and quantity of rhetorical relations. One of the most convincing arguments for the primacy of rhetorical relations is the presence of explicit triggers for some of these relations: for example, SDRT characterizes the difference between *and* and *but* by the fact that *but* signifies the presence of a *Contrast* relation. The interpretation process can then rule out any interpretation that does not support it.

Lascarides and Asher [2007] use the following example to illustrate the importance of rhetorical relations:

(3) (a) The judge asked where the defendant was.

(b) The barrister said he was in the pub drinking.

(c) The bailiff found him slumped beneath the bar.

(c′) But the bailiff found him slumped beneath the bar.

Just the presence of the word "but" completely changes the meaning of the the last statement—in (c), the man is at the pub, and in (c′), he is in the courtroom. This is because, they argue, "but" signifies a contrast in the information between the rhetorically connected clauses—(b) and (c′) in particular—and so we interpret the discourse in a way that supports the existence of a contrast.

All things considered, the SDRT approach may seem more linguistically complete and nuanced than the approach by Hobbs. However, it is comparatively impractical to use when attempting to decompose a problem like the Winograd Schema Challenge in a truly principled way. While Asher and Lascarides [2003] go to lengths to ensure that the discourse interpretation process is *decidable*, that does not mean it is computationally tractable. Furthermore, their combination of nonmonotonic reasoning with Discourse Representation Theory and discourse parsing all together make for a system that is truly complex, and for which the knowledge representation task may prove very difficult. Finally, the criteria of *quality* and *quantity* of discourse relations is somewhat fuzzy and can be hard to evaluate.

Coherence by Correlation In light of the strengths and weaknesses of these techniques, Bailey *et al.* [2015] proposed another method. It is also based on the idea of rhetorical relations, but the picture is drawn much more coarsely. While SDRT has about 14 relations, Bailey *et al.* [2015] recognize just two: positive and negative correlation. This can be (though by no means must it be) viewed as a simplification of the SDRT system: Relations such as *Explanation*, *Narration*, *Elaboration* and

Parallel roughly all involve their relata being positively correlated, while *Contrast* and perhaps *Alternation* involve their relata being negatively correlated.

They apply the framework of positive and negative correlation to the Winograd Schema Challenge, finding not only that this can help us understand the WSC, but that the WSC provides a fertile testing ground for the involvement of correlation in coherence.

2.3 The Correlation Calculus

The Correlation Calculus is a formal system proposed in Bailey *et al.* [2015] in order to facilitate the use of commonsense knowledge in justifying the answers to some anaphora resolution problems of the kind in the Winograd Schema Challenge. In particular, the *correlation formulas* of the calculus serve on one hand as the means of justification of coherence, and on the other hand as a new language by which to represent commonsense knowledge about correlation. The calculus and its use act as a small—but necessary—piece of a full story of the puzzle of discourse coherence.

This section draws heavily from Bailey *et al.* [2015].

2.3.1 Syntax

The Correlation Calculus is built on top of First-Order Logic (FOL). As such, it is defined over a first-order signature, which consists of symbols for object constants, function constants, and predicates. The language of the Correlation Calculus over a first-order signature σ consists of

(i) first-order sentences over σ, and

(ii) *correlation formulas* of the form $F \oplus G$, where F and G are first-order formulas over σ.

Intuitively, a correlation formula $F \oplus G$ with no free variables means that

a hearer will believe G to be more plausible upon hearing F, and vice versa

or

at least one of F, G, $\neg F$, $\neg G$ is known to be true.[3]

[3]The second component of this remark clarifies our treatment of the degenerate case of *complete information*, when the truth or falsity of one of the statements is known to the hearer. In the case that, say, the hearer knows F, it seems that the first component would be false. After all, the hearer cannot be convinced any more or less of the plausibility of F. But we instead treat this case in the opposite way, allowing us the *implication rule* (see Section 2.3.2).

A correlation formula with free variables can be thought of as shorthand for all of its ground instances. We may read the formula: "F is positively correlated with G." We also introduce the abbreviation

$$F \ominus G$$

which is short for

$$F \oplus \neg G$$

and may be read "F and G are negatively correlated."

2.3.2 Inference

Table 2.1 shows the rules of the calculus: implication, replacement, symmetry, negation, and substitution.[4] Each rule has a correlation formula as its conclusion and first-order sentences or correlation formulas as premises.

The symbol $\widetilde{\forall}$ denotes universal closure. The expression $F\theta$ stands for the result of applying a substitution θ of terms for variables to all free variables in F, with bound variables renamed if necessary to avoid quantifier capture.

A *derivation* from a set Γ consisting of first-order sentences and correlation formulas is a list $C_1, \ldots, C_n$ such that

- if C_i is a first-order sentence then it is entailed (in the sense of First-Order Logic) by Γ, and
- if C_i is a correlation formula then it belongs to Γ or can be derived from one or two of the formulas that precede it in the list by one of the rules of the correlation calculus.

A correlation formula C is *derivable from* Γ if there exists a derivation from Γ with C as the last formula.

The implication rule says that if a formula F implies a formula G then F and G are correlated. This is useful especially because it allows us to conclude *correlation from causation*. For example, a commonsense axiom like

$$\forall x(anchored(x) \rightarrow \neg move(x))$$

[4]Note that one of the replacement rules is redundant in light of the symmetry rule.

Implication Rule	$\frac{\widetilde{\forall}(F \rightarrow G)}{F \oplus G}$
Replacement Rule (1)	$\frac{\widetilde{\forall}(F \leftrightarrow G) \quad F \oplus H}{G \oplus H}$
Replacement Rule (2)	$\frac{\widetilde{\forall}(F \leftrightarrow G) \quad H \oplus F}{H \oplus G}$
Symmetry Rule	$\frac{F \oplus G}{G \oplus F}$
Negation Rule	$\frac{F \oplus G}{\neg F \oplus \neg G}$
Substitution Rule	$\frac{F \oplus G}{F\theta \oplus G\theta}$

Table 2.1: Inference rules of the Correlation Calculus.

could be used to derive

$$anchored(x) \oplus \neg move(x).$$

A derivation like this is used in the example of the sculpture and the shelf:

> The sculpture rolled off the shelf because it was not [anchored/level].

(See Bailey *et al.* [2015] for the initial development of this example.) This also encodes our intuition that *positive correlation* can be used where Asher and Lascarides [2003] would justify coherence with an *Explanation* relation, which is defined based on causation.

However, our approach is much coarser: not only does it use a coarser rhetorical relation, but implication in FOL is *material implication* which does not necessarily encode our intuitions about causation. We see this coarseness as an advantage in that it simplifies our reasoning and our knowledge representation language, but it comes at a price because of the semantics of material implication. On one hand, it means that positive correlation is reflexive (which makes sense), but on the other hand, it slightly complicates our intuition about correlation formulas (see Footnote 3).

2.3.3 Examples of Derivations

These derivations should give the flavor of the Correlation Calculus. They are drawn directly from Bailey *et al.* [2015].

The formula $F \oplus F$ is derivable from the empty set of formulas:

1.	$\widetilde{\forall}(F \rightarrow F)$	logically valid
2.	$F \oplus F$	by implication rule.

This shows that positive correlation is reflexive.

The formula $P(a) \oplus \exists x P(x)$ is derivable from the empty set of formulas:

1.	$P(a) \rightarrow \exists x P(x)$	logically valid
2.	$P(a) \oplus \exists x P(x)$	by implication rule.

The formula $P(a) \oplus \forall x P(x)$ is derivable from the empty set of formulas:

1.	$\forall x P(x) \rightarrow P(a)$	logically valid
2.	$\forall x P(x) \oplus P(a)$	by implication rule
3.	$P(a) \oplus \forall x P(x)$	by symmetry rule.

The formula $G \ominus F$ is derivable from $F \ominus G$:

1.	$F \oplus \neg G$	
2.	$\neg G \oplus F$	by symmetry rule
3.	$\neg\neg G \oplus \neg F$	by negation rule
4.	$\widetilde{\forall}(\neg\neg G \leftrightarrow G)$	logically valid
5.	$G \oplus \neg F$	by replacement rule.

This derivation shows that negative correlation, like positive correlation, is symmetric.

For any first-order formulas F and G, the correlation formula $F \oplus G$ is derivable from $\widetilde{\forall} F$:

1.	$\widetilde{\forall} F$	
2.	$\widetilde{\forall}(G \rightarrow F)$	entailed by 1
3.	$G \oplus F$	by implication rule
4.	$F \oplus G$	by symmetry rule.

It is derivable from $\widetilde{\forall}\neg F$ as well:

1.	$\widetilde{\forall}\neg F$	
2.	$\widetilde{\forall}(F \rightarrow G)$	entailed by 1
3.	$F \oplus G$	by implication rule.

These last two derivations bear out the second component of the intuitive understanding of correlation formulas given in Section 2.3.1. Because of the implication rule, any formula known to be true or false may appear on either side of a correlation formula.

2.3.4 Application to the WSC

Consider the sentence from Schema 2 on Davis's list:

The trophy doesn't fit into the brown suitcase because it's too small. What is too small?

We will justify the correctness of the answer *the suitcase* as follows. The phrase *the trophy doesn't fit into the brown suitcase* can be represented by the sentence $\neg \mathit{fit\_into}(T, S)$, with the presuppositions

1. $\mathit{trophy}(T)$
2. $\mathit{suitcase}(S)$
3. $\mathit{brown}(S)$.

The phrase *the suitcase is too small* can be represented by the sentence $\mathit{small}(S)$. We will show that the correlation formula

$$\neg \mathit{fit\_into}(T, S) \oplus \mathit{small}(S) \tag{2.1}$$

can be derived in the Correlation Calculus from the presuppositions 1–3 in combination with the following commonsense facts, which are assumed to be known to the hearer:

4. $\forall x(\mathit{suitcase}(x) \rightarrow \mathit{physical\_object}(x))$
5. $\forall x(\mathit{physical\_object}(x) \rightarrow (\mathit{small}(x) \leftrightarrow \neg \mathit{large}(x)))$
6. $\mathit{fit\_into}(x, y) \oplus \mathit{large}(y)$.

The derivation continues as follows:

7. $\neg \mathit{large}(S) \leftrightarrow \mathit{small}(S)$	entailed by 2, 4, and 5
8. $\mathit{fit\_into}(T, S) \oplus \mathit{large}(S)$	by substitution from 6
9. $\neg \mathit{fit\_into}(T, S) \oplus \neg \mathit{large}(S)$	by negation rule
10. $\neg \mathit{fit\_into}(T, S) \oplus \mathit{small}(S)$	by replacement rule.

In Axiom 6 above, x and y are, intuitively, physical objects, and moreover y is a container. Our formulation may seem too strong because it does not incorporate these assumptions about the values of x and y. Note, however, that when x and y are not objects of appropriate types, the corresponding instance of Axiom 6 holds because at least one of the conditions $\mathit{fit\_into}(x, y)$, $\mathit{large}(y)$ is known to be false.

The derivability of (2.1) shows that

$$\neg \textit{fit\_into}(T, S) \text{ is correlated with } \textit{small(S)}$$
$$\text{or}$$
$$\text{at least one of } \textit{fit\_into}(T, S),\ \textit{small(S)}$$
$$\text{is known to be true or false.}$$

Since the formulas $\textit{fit\_into}(T, S)$ and *small(S)* are not known to be true or false, we can conclude that the sentences $\neg\textit{fit\_into}(T, S)$ and *small(S)* are indeed correlated.

Consider now the same example with the special word *small* replaced by the alternate word *big*:

The trophy doesn't fit into the brown suitcase because it's too big.

What is too big? The correctness of the answer *the trophy* can be justified in a similar way, with Axiom 6 replaced by the axiom

$$\textit{fit\_into}(x, y) \oplus \textit{small}(x).$$

2.3.5 Obtaining Axioms

The Correlation Calculus is not itself a solution to the Winograd Schema Challenge. It is only a small piece of a much larger story. We see the Winograd Schema Challenge as involving three main problems:

(i) Determining the nature of the requisite knowledge,

(ii) Creating a database of this knowledge, and

(iii) Applying this knowledge to anaphora resolution problems.

The Correlation Calculus is an approach for the third. However, it has implications for the first, in that it gives us a new representation language for certain kinds of knowledge.

The questions that remain are how to obtain this knowledge and what it should look like. One criticism of the approach in Bailey *et al.* [2015] is that the axioms we use in each example are tailored to the problem at hand. This is unfortunately

necessary when demonstrating a new approach to problem (iii) independently of an approach to (ii), as we are doing, because databases of correlation formulas don't exist. (Another argument is that a project to gather this kind of knowledge is only justified once a use for it has been established.) Regardless of this issue, in a real system that uses the Correlation Calculus to solve WSC problems, we would not expect such tailor-made axioms to be available. Rather, what we hope is that the approach we outline can be used with axioms that are few or general enough to write down by hand or extract from data—i.e., the proofs from these axioms might be longer or more involved than in the examples we present. The only way to show this conclusively would be to do it, and that would be a project in itself.

Various groups have suggested ways to obtain from data the knowledge necessary for the WSC; it is possible that some of these approaches may be modified and adapted to the Correlation Calculus. See Bailey *et al.* [2015], Levesque [2014], Sharma [2014], Budukh [2013], and Sharma *et al.* [2015] for discussions.

2.4 Discourse Representation Theory

While First-Order Logic can express a great deal of truth-conditional meaning, it falls short as a meaning representation language when it comes to the dynamics of discourse. Phenomena like anaphora, presupposition, and tense and aspect are difficult to account for using FOL. Discourse Representation Theory (DRT) was developed to ameliorate these difficulties (see Kamp and Reyle [1993]). It is also seen as a *representationalist* language, that is, its basic units (called Discourse Representation Structures, or DRSs) in some sense represent the mental state of a participant of the discourse.

Since DRT was originally developed as a treatment of anaphora, it is particularly suited to our purposes in the Winograd Schema Challenge. Furthermore, if we wish to approach the problem of anaphora resolution in the Winograd Schema Challenge in a linguistically credible way—as the "explanation of a phenomenon" desired by Levesque [2014]—we must develop our use of correlation in a framework that to the best of our knowledge represents the phenomenon faithfully. This is the capacity in which DRT will prove useful for our approach to the WSC.

DRT is an extensive system and a complete treatment is not possible even in a large volume, let alone this document. For this reason we will gloss over most of the details irrelevant to our use of the system. Furthermore, while we will be able

to provide examples to illustrate some of the motivations behind the structure of DRT, fully justifying the system is something we leave to the extensive literature on the topic. On the whole, this section will follow the development of DRT in Kamp and Reyle [1993].

2.4.1 Motivation for DRT

Discourse Representation Theory was originally motivated by two main problems that remained difficult under traditional FOL-based models of linguistic meaning. One is what is often called *donkey anaphora*, which is a phenomenon involving anaphoric reference that seems to violate the scoping restrictions of First-Order Logic. The other is a proper treatment of tense and aspect, particularly in light of the French *Passé Simple* and *Imparfait*, two distinct past tenses. The first is easier to explain, and we will illustrate it briefly here.

Donkey anaphora is the kind of reference that the word *it* makes in the following sentence:

(4) If Pedro owns a donkey, then he beats it.

The problem is this: consider the following FOL representation of "Pedro owns a donkey."

$$\exists a \exists b(Pedro(a) \wedge donkey(b) \wedge own(a, b)).$$

By the "if–then" structure of 4, we would like this to be the antecedent of the conditional being expressed. But there we run into a problem:

(5) $\exists a \exists b(Pedro(a) \wedge donkey(b) \wedge own(a, b)) \rightarrow beat(a, b)$

The result is not what we want: a and b are free in the consequent of (5), so they don't refer to Pedro or the donkey he owns in any meaningful way. A first-order formula that better approximates the apparent truth conditions of this sentence might look like:

(6) $\exists a(Pedro(a) \wedge (\forall b(donkey(b) \wedge own(a, b)) \rightarrow beat(a, b)))$.[5]

[5]Under this interpretation, sentence (4) has the same truth conditions as *Pedro beats every donkey he owns.* This interpretation is accepted in the literature, but in general a *weaker* reading of *every* is possible, where for example

Every man who has a nice suit will wear it to church tomorrow.

means that every such man will wear *some* such suit. The felicity of examples like this is debatable, but other examples involving generalized quantifiers like *most* seem to indicate that both the weak and strong readings are possible (Kanazawa [1994]).

But it's not straightforward how we would derive such a formula from the syntactic structure of (4). The takeaway from this example is that variable scoping in natural language operates very differently from variable scoping in FOL, and that makes it difficult to see how the meanings of natural language sentences would compose together into a discourse. That's also not to mention other subtleties about this example: Why does Pedro appear under an existential but the donkey a universal? In what range of time is Pedro supposed to own the donkeys, and at what points of time is he supposed to be beating them? Discourse Representation Theory was developed to answer these kinds of questions.

The core departure DRT makes from a FOL-based approach is that it recognizes discourse as *dynamic*. As a discourse progresses, new entities and ideas are introduced, and these may serve as anchors for anaphoric reference, whether it means providing antecedents for pronouns or establishing the progression of time through a narrative. How exactly it does this will be made clear in the following sections.

2.4.2 Formal Development

The language of DRT is the language of Discourse Representation Structures (DRSs) and DRS-Conditions. It has a model-theoretic semantics and a sound and complete calculus for its first-order fragment (see Kamp and Reyle [1996]). In this section we will begin with overview of the basic syntax and semantics of DRT, as well as a brief discussion of event semantics, each to the extent to which it will be important in our application to establishing discourse coherence.

Syntax We begin with a *vocabulary*, which consists of:

- A set of names for individuals: $John, Obama, \ldots$ [6]
- A set of predicate constants of various arities: $donkey^1, own^2, beat^2, \ldots$ etc.

The syntax of DRSs is defined with respect to a vocabulary V and a set R of *discourse referents*, which form a second alphabet disjoint from V. The *DRS language* is defined with respect to V and R, and its definition is co-recursive between DRSs and DRS-conditions. A DRS K confined to V and R contains:

[6] Depending on the particular way proper names are formalized, these are used differently. In our approach, they will appear in the same position as predicates—e.g., $Obama(x)$—whereas for example, the Boxer system (Bos [2008]) will use them in DRS-conditions like $named(x, obama)$. In either case, the meaning is the same.

- a subset of R, called the *universe* of the DRS, and
- a set of *DRS-conditions* confined to V and R.

DRS-conditions are of two types: *simple* and *complex*. A simple DRS-condition confined to V and R is one of the following:

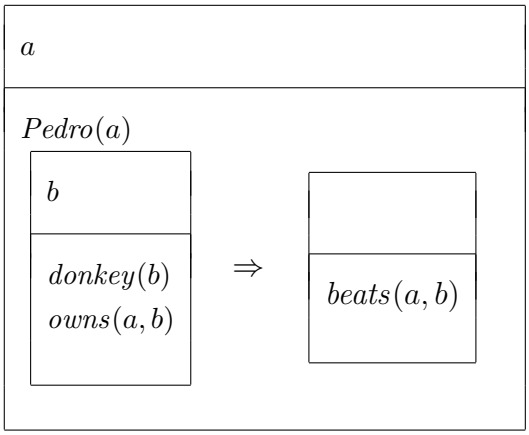
- $x = y$, where $x, y \in R$,
- $\pi(x)$, where $x \in R$ and π is a name from V,
- $P(\overline{x})$, where P is an n-ary predicate from V and $\overline{x}$ is an n-tuple of elements of R.

Proper names are interpreted differently from predicates, in that they correspond to exactly one individual. A complex DRS-condition is one of several constructions on DRSs. Given two DRSs K_1 and K_2 confined to V and R, a complex DRS-condition confined to V and R is

- $\neg K_1$, the *negation* of K_1,
- $K_1 \Rightarrow K_2$, the *conditional* construction, and
- $Q_x(K_1, K_2)$, where Q is some quantifier and x is in the universe of K_1—called a *duplex condition*.

In a duplex condition as above, K_1 is called the *restrictor* and K_2 is called the *nuclear scope*. The first-order fragment of DRT can be fully expressed without the use of this kind of condition. More generally, DRT often includes treatments of generalized quantifiers like *most*, but that is not relevant to our current analysis.

For readability, DRSs are usually written in a "box notation." For example, the DRS with the right truth conditions for sentence (4) would look something like



Each box is a DRS. The section at the top of each box is its universe, and its DRS-conditions are listed in its body.

The list of DRS-conditions is to be understood conjunctively, i.e., a DRS asserts that all of its DRS-conditions hold. Intuitively, we may think of a discourse referent as a bound variable, bound where it appears in the universe of its DRS. It furthermore is understood to be existentially quantified. But then it might seem strange that a and b both appear in the consequent sub-DRS above, though they are not in its universe. This is where the difference between DRT and FOL begins to show. The scope of discourse referents is determined by the *accessibility* relation between DRSs.

Accessibility is the smallest relation satisfying the following conditions:

- Accesibility is reflexive and transitive.
- If K_2 is part of a DRS-condition of K_1, K_1 is accessible to K_2.
- In $K_1 \Rightarrow K_2$, K_1 is accessible to K_2.
- In $Q_x(K_1, K_2)$, K_1 is accessible to K_2.

The use of a discourse referent in a simple DRS-condition of K is *bound* if it appears in the universe of a DRS that is accessible to K. Otherwise, it is *free*. We call K a *proper* DRS if no discourse referents are free in K. This is analogous to being a closed formula, in that only a proper DRS has a semantics that assigns it a truth value.

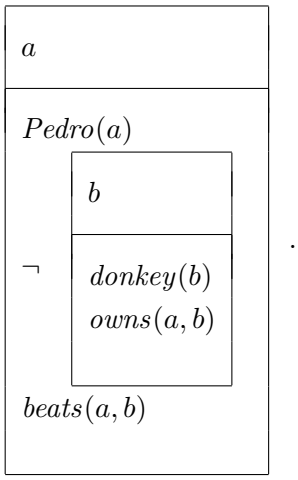
There are a few other nuances of accessibility, but this is the major picture. Conditionals tell the main point of the story: if the antecedent of a conditional supposes the existence of some individual, the consequent has license to assume the existence of that individual and refer back to it.[7] The accessibility relation provides a way to explicitly discuss the scoping and binding of discourse referents in DRT, and compare them to the constraints on anaphora and binding in natural language. Take, for example, the discourse

(7) Pedro does not have a donkey. #He beats it.[8]

[7] This point may make it seem that in DRT, the conditional and universal quantifier may coincide. By some analyses, including the semantics we will give, this is true. This fact and the reasoning supporting it exhibit a parallel with the dependent product type (or Π-type) from dependent type theory, indicating a connection to the seemingly very different approach of type-theoretic semantics taken by Ranta [1995] and Cooper [2005].

[8] Here the # mark is used to mark the sentence as semantically infelicitous.

We may try to translate it to the DRS



.

But here, b in $beats(a, b)$ is free, in very much the same sense that *it* in *he beats it* is meaningless. The accessibility relation, in some sense, is meant to model the natural, logical constraints faced by binding and anaphoric reference in discourse.

The accessibility relation is not just a matter of syntax: discourse referents appearing outside of the DRS in whose universe they appear also need a well-defined semantics. In fact, this issue is where the formal model-theoretic semantics of DRT differs most from that of FOL. It uses *embedding functions*, which assign discourse referents to elements in the universe of the model, and pass them between DRSs in a way that corresponds exactly to the accessibility relation to ensure that every bound discourse referent will take a value when it is interpreted.

However, for a proper, finite DRS, there is another way to view their semantics.[9] It turns out that the finite first-order fragment of DRT is in a certain sense isomorphic to FOL: There is a translation between proper DRSs and FOL sentences that preserves validity and entailment both ways. The exact semantics of DRT is not important in this work, but this translation *is*, and it also provides us an understanding of the semantics. So instead of fully developing the semantics of DRT, we will continue here with the translation to FOL.

[9] A DRS is finite if its universe is finite, it has finitely many DRS-conditions, and all of its subordinate DRSs are finite. This is not part of the standard definition of DRSs, but this is the condition under which the correspondence to (finitary) FOL holds. For natural language discourse there is little reason to think we would need infinite DRSs.

2.4.3 Translating from DRT to FOL

Let K be a proper finite DRS over V and R, with universe $\{x_1, \ldots, x_n\} \subseteq R$ and DRS-conditions $\{\gamma_1, \ldots, \gamma_m\}$. Then the *translation* $[[K]]$ of K is a first-order formula over a signature with the predicate constants from V, the names from V as object constants, and no function constants. It reads

$$\exists x_1 \ldots \exists x_n \bigwedge_{i=1}^{m} [[\gamma_i]].$$

For a DRS-condition γ, its *translation* $[[\gamma]]$ is defined as follows.

- If γ is an instance of a linguistic predicate-argument structure, such as **a hit b**, then $[[\gamma]]$ will be its corresponding first-order formula under the assumed correspondence between predicate-argument structures and predicates in the signature, e.g., $hit(a, b)$.
- If γ is $\pi(a)$, where π is a name in V and $a \in R$, then $[[\gamma]]$ will be the formula $a = \pi$.[10]
- If γ is some other simple DRS-condition, then $[[\gamma]]$ is simply the string γ, which is a valid first-order formula.
- If γ is of the form $\neg K'$, $[[\gamma]]$ is the formula $\neg[[K']]$.
- If γ is of the form $K_1 \Rightarrow K_2$, let $\{y_1, \ldots, y_j\}$ be the universe of K_1 and $\{\delta_1, \ldots, \delta_k\}$ be its DRS-conditions. Then $[[\gamma]]$ is

$$\forall y_1 \ldots \forall y_j ((\bigwedge_{i=1}^{k} [[\delta_i]]) \rightarrow [[K_2]]).$$

2.4.4 The Syntax-Semantics Interface

An theory of the semantics of natural language would not be very useful without an account of how logical forms may be derived from an utterance. There are multiple ways to construct DRSs from discourse (Blackburn and Bos [1999]), but we will focus on the top-down method outlined by Kamp and Reyle [1993], which constructs a DRS from one or more syntax trees by the application of *construction rules* on

[10] In this formula, a will be a bound variable. It will have been bound by an existential quantifier introduced in the translation of the DRS in whose universe a appeared.

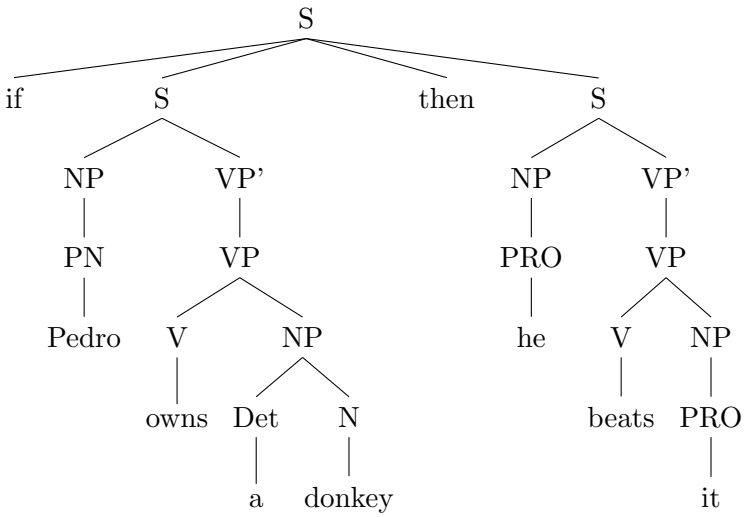
instances of *triggering configurations* in the trees. Their approach also makes use of a certain form of *Generalized Phrase Structure Grammar* (Gazdar *et al.* [1985]), and we will use parse trees from this grammar, but the particular choice of syntactic formalism is immaterial. Kamp and Reyle [1993] said it best:

> [W]e have stayed aloof from the question whether our rules capture the deeper syntactic regularities which a self-respecting syntactician would see it as his primary duty to discover. We believe, however, that such expedience is fairly harmless. For all existing experience with the semantic theory we present here suggests that it can be easily adapted to fit a theory of syntactic form which does justice to the issues that make syntax an important and fascinating subject in its own right.

To illustrate the process, we will use the discourse

(8) If Pedro owns a donkey, then he beats it.

The (greatly simplified) syntactic analysis we will use is as follows:



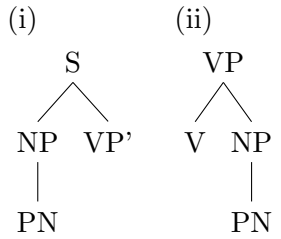
The DRS Construction Algorithm incrementally converts a discourse like the above into its DRS representation by the repeated application of *construction rules* to the sentences in sequence.

Each construction rule consists of a set of *triggering configurations* and a procedure. The triggering configurations are schemas of parse tree fragments, and the procedure is a series of operations that change the DRS and the tree, using information from an instance of one of the triggering configurations.

To analyze item 8, we will need several construction rules.

(9) **Construction Rule CR.PN** (Proper Names).

Triggering Configurations.



Procedure.

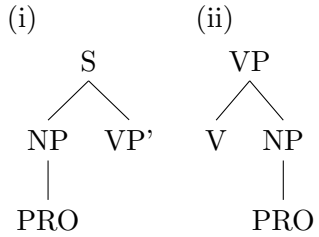
(i) Introduce a new discourse referent into the universe of the outermost DRS. Call it u.

(ii) Let *Name* be the proper name referenced in the triggering configuration. Add the condition $Name(u)$ to the condition set of the outermost DRS.

(iii) In the tree with the triggering configuration, replace the subtree rooted at the NP node in the triggering configuration with u.

(10) **Construction Rule CR.PRO** (Pronouns).

Triggering Configurations.



Procedure (non-deterministic).

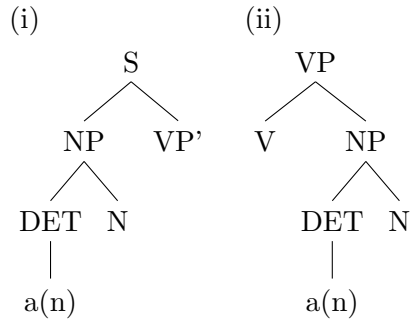
(i) Introduce a new discourse referent into the universe. Call it u.

(ii) In the tree with the triggering configuration, replace the subtree rooted at the NP node in the triggering configuration with u.

(iii) Add a new condition of the form $u = \alpha$ where α is some non-deterministically chosen accessible discourse referent.

(11) **Construction Rule CR.ID** (Indefinites).

Triggering Configurations.



Procedure.

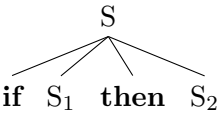
(i) Introduce a new discourse referent into the universe. Call it u.

(ii) In the tree with the triggering configuration, substitute u for

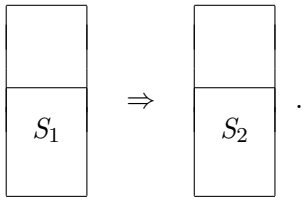
(iii) Add a new condition of the form $[N](u)$,[11] where $[N]$ is the noun appearing under the N node.

(12) **Construction Rule CR.COND** (Conditionals).

Triggering Configurations.



Procedure. Replace the condition containing the triggering configuration by



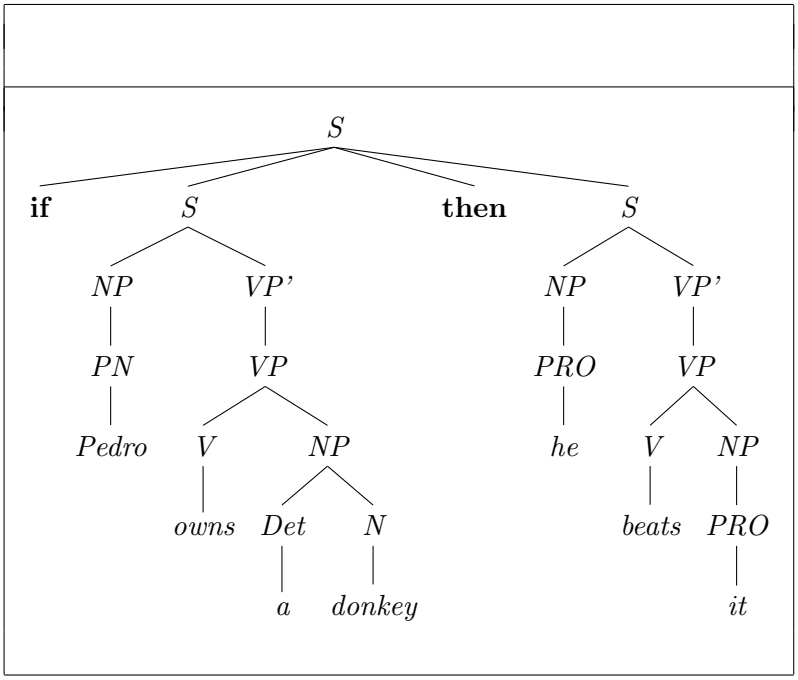
.

In the intermediate stages of DRS construction, the object we are working with may not strictly be a DRS, in the sense that it contains some reducible conditions

[11]This is slightly different from the rule in Kamp and Reyle [1993], which sets up a new condition to be picked up later by the *lexical insertion* rule. This complication is used to deal with relative clauses, which are not relevant to our analysis, so we will slightly simplify it.

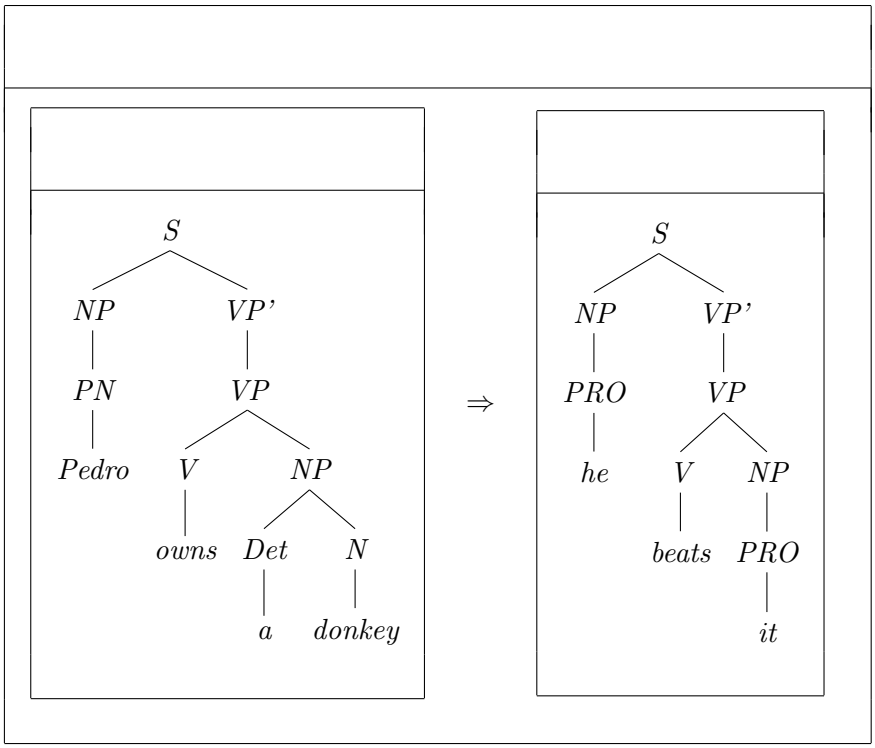
that do not have an exact semantics. We will gloss over this fact and call it a DRS anyway.

Using these rules, we process the discourse (8) as follows. We begin by adding the syntactic analysis to the DRS:

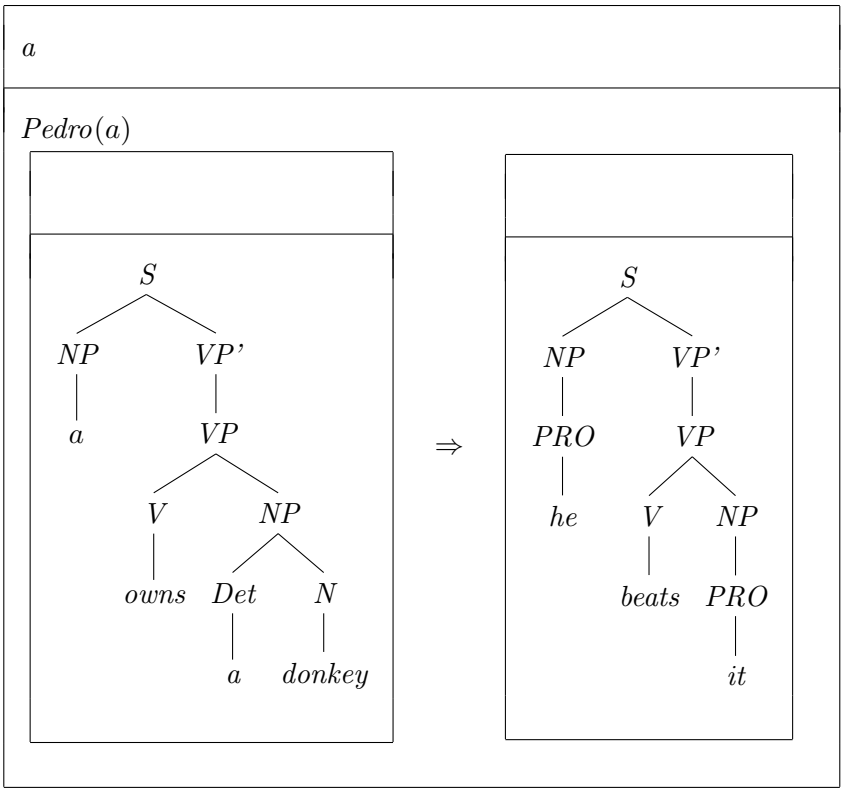


.

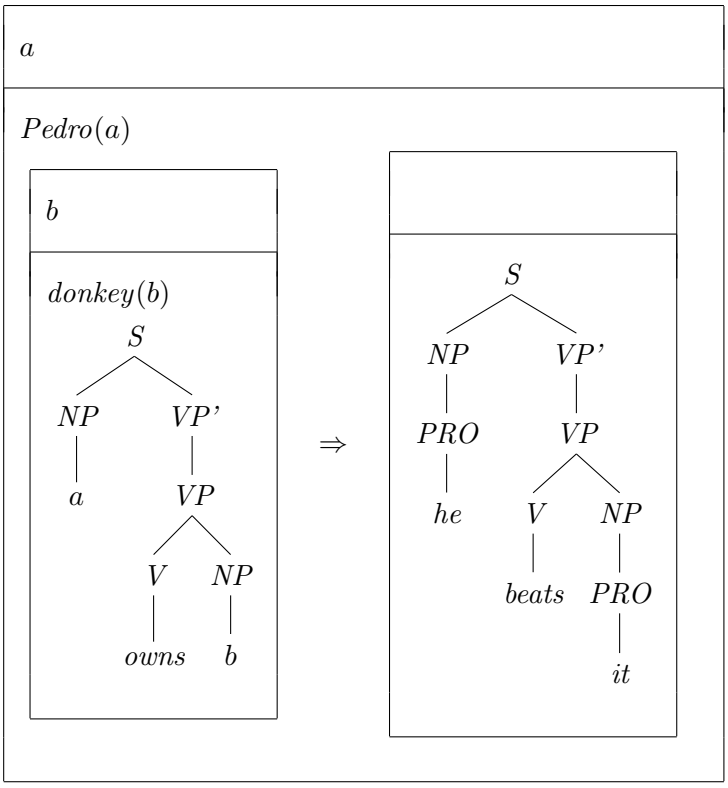
First we apply **CR.COND** to get



which establishes the conditional structure of the proposition. Next, **CR.PN** gives us



where the discourse referent for *Pedro* was put in the outer DRS because it is a proper name. Next, an application of **CR.ID** yields



so the phrase *a donkey* is interpreted as introducing the existence of a donkey under the scope of the original *if*.[12]

The last condition in the antecedent DRS—in the form of a syntactic tree—is *irreducible*, and must be interpretable as something like an atomic formula in the vocabulary underlying our DRS language. To facilitate this, in DRT, we assume a faithful embedding from predicate-argument structures of verbs into the set of predicates in the vocabulary. We then *identify* the irreducible syntactic trees corresponding to instances of these predicate-argument structures with atomic formulas according to that embedding. So it would be equivalent—and perhaps more transparent—to write the condition as

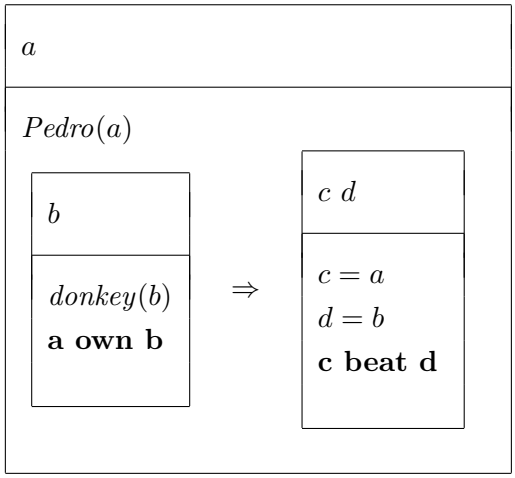
$$own(a, b).$$

[12]DRT acknowledges that there may also be scope ambiguities with indefinites, i.e., in general, an indefinite may be able to scope out of a negation or conditional. This may be modeled in the syntactic analysis, as in the *quantifier raising* approach of Heim and Kratzer [1998]. Another way this is modeled is in the *singular* interpretation of indefinites, where the indefinite is made to scope to the outermost DRS (Kamp and Reyle [1993]).

But for brevity (and clarity when we introduce event semantics in Section 2.4.5) we will follow Kamp and Reyle [1993] and abbreviate the condition simply as **a own b**.

Finally, two applications of **CR.PRO** yield the final result (using the new shorthand):

(13)

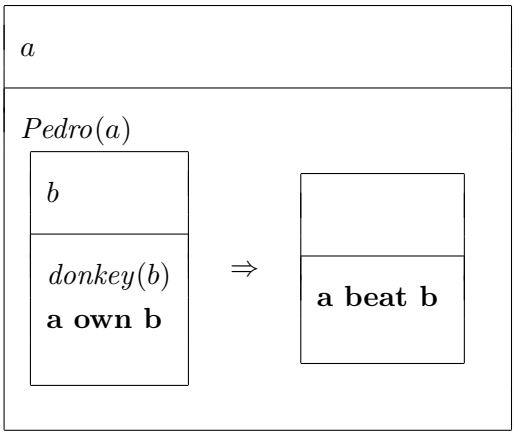


.

In principle these steps are nondeterministic and the new referents c and d could have been chosen to equate with different antecedents. It happens that, due to binding and agreement constraints in English, this particular discourse is completely unambigious—the antecedents can be decided on those grounds alone. But in general, and particularly in the Winograd Schema Challenge, this information will not be enough. The question of how to make this choice is left unanswered in the construction algorithm. Indeed, the question of how to choose the antecedent is the primary one of this work, and of the Winograd Schema Challenge. But since we are just presenting the construction algorithm in this section, we've gone the route of angelic nondeterminism.

There are many more construction rules, dealing with more complicated phenomena like quantification, plurals, tense and aspect. More yet would be necessary for a broad-coverage system. The system Boxer uses a different DRS construction method (Bos [2008]). But the top-down system we present, which is used in Kamp and Reyle [1993], is sufficient for our purposes with the Correlation Calculus and the Winograd Schema Challenge.

To understand our results a little better: it's easy to see from the translation to FOL that the DRS (13) is equivalent to



which translates to FOL as

$$\exists a(Pedro(a) \wedge \forall b((donkey(b) \wedge own(a,b)) \rightarrow beat(a,b))).$$

This is exactly the formula (6) given in the motivation for DRT.

We now present the construction algorithm in full.

(14) **Algorithm** (DRS Construction).

Input. Syntactic analyses for one or more sentences that, in sequence, comprise a discourse.

Output. A proper Discourse Representation Structure serving as a logical form of the input discourse's syntactic analysis.

Procedure (nondeterministic).

(i) Begin with an empty DRS.

(ii) Place the syntactic analysis of the first unprocessed sentence into the DRS.

(iii) Apply construction rules to reducible conditions in the DRS until all of the conditions are irreducible.

- If there are multiple triggering configurations in a reducible condition, apply the rule to the highest one first. If there is a tie, nondeterministically break it.
- If there are multiple reducible conditions, nondeterministically choose one to reduce.

Note that the application of construction rules also may have nondeterministic results.

(iv) Repeat from (ii) until there are no more sentences to be processed.

Note that the result of step (iii) will always strictly be a DRS in the sense that all of its conditions are irreducible. So when this algorithm terminates, the result will be a DRS, under the assumption that the set of construction rules can always reduce the syntactic analyses of the discourse to DRS-conditions and (linguistic) predicate-argument structures.

2.4.5 Event Semantics

The account above could answer the question of donkey anaphora, but it left tense and aspect, and temporal reference generally, untouched. This side of DRT draws on the notion of *event semantics*, which will play a role in relating the Correlation Calculus to DRT.

At the core of our approach is acknowledging that positive and negative correlation, as rhetorical relations, may be expected to hold between certain statements in a coherent discourse. In order to extract correlation formulas from DRS representations, we need some way of representing, on a very basic level, the rhetorical structure of the discourse in question. The way we will manage to do that is by acknowledging that clauses fundamentally refer to *events* and *states*. This will facilitate the use of correlation within DRT, but it may have beneficial side-effects in resolving temporal reference as well (see Chapter 5).

Event semantics begins with Davidson [1967], who suggested that events should be *reified* in FOL descriptions of natural language meaning. Instead of representing *a hit b* as the proposition

$$hit(a, b),$$

it might be

$$\exists e(hit(e, a, b)).$$

This has the advantage that other aspects of the event—such as the time at which it occurred—may be included in the formula or not:

$$\exists t \exists e(hit(e, a, b) \wedge time(e, t) \wedge midnight(t))$$

is a simple example of how temporal information may be added to the above formula. In this case, Davidson argued, the deduction that *a hit b* from *a hit b at midnight* can be done by the inference rules of predicate logic. How exactly one proceeds

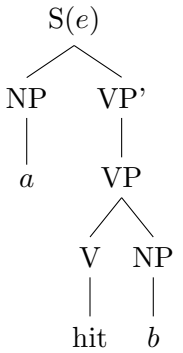
once this reification is done is a matter that different groups have taken in different directions. The "ontological promiscuity" approach of Hobbs [1985] follows closely along these lines, while the "neo-Davidsonian" approach uses semantic roles, as in:

$$\exists e(hit(e) \wedge agent(e, a) \wedge patient(e, b)).$$

Of these options, we will adopt the simplest version, which is the approach taken in Kamp and Reyle [1993], where a predication like *a hit b* will be represented by

$$hit(e, a, b),$$

where e, a, and b will be discourse referents in some accessible universe (determined during the construction of the DRS). However, this (the 3-place *hit* predicate) will only be its representation on the level of the *model*, i.e., the semantic interpretation. The irreducible condition will actually look like



which will be abbreviated as

e: **a hit b**

retaining (on a syntactic, though perhaps not semantic level) the special status of the event-denoting discourse referent. Note also that in the irreducible condition we have the annotation of e on the S-node. This is a result of changes made to the construction algorithm upon the introduction of event semantics (which do not affect our analysis). We refer the reader to Kamp and Reyle [1993] for details.

Chapter 3

Semantics of the Correlation Calculus

We can make our intuitions about correlation precise with a probabilistic semantics of the Correlation Calculus. With it, we have a soundness result for the calculus as well as a way to understand correlation formulas in terms of probabilistic correlation.

Recall that a correlation formula $F \oplus G$ is understood to mean that a hearer will judge F more plausible upon hearing G, and vice versa—or the hearer already knows that F or G is true or false. We make this notion precise by understanding plausibility and knowledge in terms of *probability*, and the "upon hearing" stipulation as *conditionalizing*. In this way, correlation formulas can be thought of as something like statements about a probabilistic belief state over worlds, where the analogy for first-order formulas would be statements about a world.

In this section, we assume that the underlying signature is finite and contains at least one object constant but no function symbols of arity greater than zero. Under these assumptions, the set $\mathbf{I}$ of Herbrand interpretations of the signature is finite.

3.1 Worldviews and Satisfaction

A *worldview* is a discrete probability distribution over $\mathbf{I}$ (a function assigning a value in $[0, 1]$ to each interpretation, so that the sum over $\mathbf{I}$ is 1). Every first-order sentence defines an event—the set of its models—so we may talk about its probability with

respect to a worldview D:

$$P(F) = \sum_{I \in \mathbf{I} \,:\, I \models F} D(I).$$

Similarly, we can talk about the probability of a set Γ of first-order sentences with respect to D:

$$P(\Gamma) = \sum_{I \in \mathbf{I} \,:\, I \models \Gamma} D(I).$$

A variable is *free* in $F \oplus G$ if it is free in F or in G. A *correlation sentence* is a correlation formula with no free variables. *Satisfaction* is defined as follows: for any worldview D,

(i) D satisfies a first-order sentence F if

$$P(F) = 1,$$

(ii) D satisfies a correlation sentence $F \oplus G$ if the inequality

$$P(F \wedge G) > P(F)P(G)$$

holds, or D satisfies at least one of the sentences

$$F, \neg F, G, \neg G,$$

(iii) D satisfies a correlation formula $F \oplus G$ with free variables if D satisfies $F\theta \oplus G\theta$ for every substitution θ that maps the free variables to object constants,

(iv) D satisfies a set consisting of correlation formulas and first-order sentences if it satisfies all elements of the set.

Case (i) of the definition above is a generalization of the usual definition of satisfaction for first-order sentences: for any worldview D that assigns probability 1 to an interpretation I, D satisfies a formula F if and only if I satisfies F.

Furthermore, if a worldview satisfies a set Γ of first-order sentences then $P(\Gamma) = 1$. Indeed, the set of interpretations satisfying Γ is the intersection of the events $\{I \in \mathbf{I} \,:\, I \models F\}$ for all F in Γ. The probability of each of these events is 1, and there are finitely many of them.

In the case when $P(G) > 0$, the inequality in clause (ii) of the definition of satisfaction can be rewritten in terms of conditional probabilities; D satisfies $F \oplus G$ if

$$P(F|G) > P(F).$$

Let C be a correlation formula or first-order sentence. A set Γ of first-order sentences and correlation formulas *entails* C if every worldview that satisfies Γ satisfies C. In the special case when C and all elements of Γ are first-order sentences, this definition is equivalent to the usual definition in first-order logic. Indeed, it is clear that if every worldview satisfying Γ satisfies C then every interpretation satisfying Γ satisfies C. In the other direction, assume that every interpretation satisfying Γ satisfies C. Then $P(C) \geq P(\Gamma)$. If a worldview D satisfies Γ then $P(\Gamma) = 1$, so that $P(C) = 1$. Thus Γ entails C in the sense of the definition above.

3.2 Soundness

Consider an inference rule such that in its instances

$$\frac{C_1 \ \cdots \ C_n}{C_{n+1}} \tag{3.1}$$

$C_1, \ldots, C_{n+1}$ are either correlation formulas or first-order sentences. Such a rule is *sound* if for each of its instances (3.1) C_{n+1} is entailed by $C_1, \ldots, C_n$.

Soundness Theorem. *All rules of the correlation calculus are sound.*

Corollary. *If C is derivable from Γ in the correlation calculus then Γ entails C.*

Proof (Soundness).

The implication rule. Consider first the special case when F and G are sentences:

$$\frac{F \to G}{F \oplus G}.$$

If either $P(F) = 0$ or $P(G) = 1$ then the fact that D satisfies the conclusion is immediate. Otherwise, let D be a worldview that satisfies the premise. Then for any interpretation I to which D assigns nonzero probability, $I \models F \to G$. So for any such interpretation, if $I \models F$ then $I \models G$. Consequently

$$P(F \wedge G) = P(F).$$

On the other hand, since $P(F) > 0$ and $P(G) < 1$,

$$P(F) > P(F)P(G).$$

Consequently $P(F \wedge G) > P(F)P(G)$, so D satisfies the conclusion.

Consider now an instance of the implication rule in which F and G may contain free variables:

$$\frac{\widetilde{\forall}(F \rightarrow G)}{F \oplus G}. \tag{3.2}$$

Let D be a worldview that satisfies the premise. We need to show that for any substitution θ that maps the free variables to object constants, D satisfies $F\theta \oplus G\theta$. Given any such θ, consider the following instance of the implication rule:

$$\frac{F\theta \rightarrow G\theta}{F\theta \oplus G\theta}. \tag{3.3}$$

Since D satisfies the premise of (3.2), it satisfies the premise of (3.3). Since (3.3) is covered by the special case discussed earlier, it follows that D satisfies $F\theta \oplus G\theta$.

The replacement rule. Consider first an instance of the replacement rule where F, G, H are sentences:

$$\frac{F \leftrightarrow G \qquad F \oplus H}{G \oplus H}.$$

Let D be a worldview that satisfies both premises. Since D satisfies the first premise, it assigns probability 0 to interpretations satisfying $F \wedge \neg G$ and to interpretations satisfying $\neg F \wedge G$. Consequently,

$$\begin{aligned} P(F) &= P(F \wedge G) + P(F \wedge \neg G) \\ &= P(F \wedge G) \\ &= P(F \wedge G) + P(\neg F \wedge G) \\ &= P(G), \end{aligned}$$

so that

$$P(F) = P(G). \tag{3.4}$$

Similarly,

$$P(F \wedge H) = P(G \wedge H). \tag{3.5}$$

Since D satisfies the second premise, two cases are possible:

$$P(F \wedge H) > P(F)P(H)$$

or one of the probabilities $P(F), P(H)$ is 0 or 1. In the first case, by (3.4) and (3.5),

$$P(G \wedge H) = P(F \wedge H) > P(F)P(H) = P(G)P(H).$$

In the second case, in view of (3.4), one of the probabilities $P(G), P(H)$ is 0 or 1. In either case, D satisfies the conclusion.

The general case follows as in the proof for the implication rule. The proof for the other replacement rule is analogous.

The symmetry rule. Obvious.

The negation rule. Consider first an instance of the negation rule where F, G are sentences:

$$\frac{F \oplus G}{\neg F \oplus \neg G}.$$

Let D be a worldview satisfying the premise, so that either

$$P(F \wedge G) > P(F)P(G)$$

or one of the probabilities $P(F)$, $P(G)$ is 0 or 1. In the first case,

$$\begin{aligned} P(\neg F \wedge \neg G) &= P(\neg(F \vee G)) \\ &= 1 - P(F \vee G) \\ &= 1 - P(F) - P(G) + P(F \wedge G) \\ &> 1 - P(F) - P(G) + P(F)P(G) \\ &= (1 - P(F))(1 - P(G)) \\ &= P(\neg F)P(\neg G), \end{aligned}$$

so D satisfies the conclusion. In the second case, one of the probabilities $P(\neg F)$, $P(\neg G)$ is 0 or 1, and consequently D satisfies the conclusion.

The general case follows as above.

The substitution rule. Consider an instance of the substitution rule

$$\frac{F \oplus G}{F\theta \oplus G\theta},$$

and let D be a worldview satisfying the premise. For any substitution θ' that maps the free variables of $F\theta$, $G\theta$ to object constants, D satisfies

$$(F\theta)\theta' \oplus (G\theta)\theta'$$

because $(F\theta)\theta' = F(\theta\theta')$, $(G\theta)\theta' = G(\theta\theta')$. $\square$

3.3 Uses of the Semantics

As an example of the use of the probabilistic semantics, consider a first-order signature allowing only two distinct ground atoms p, q. We will show that the correlation formula $p \oplus q$ is not derivable from the empty set. Consider the worldview that assigns the same value $\frac{1}{4}$ to each of the 4 interpretations of this signature. This worldview does not satisfy $p \oplus q$, because

$$P(p \wedge q) = P(p) \cdot P(q).$$

Indeed, $P(p \wedge q) = \frac{1}{4}$ and $P(p) = P(q) = \frac{1}{2}$.

Here is another use of the semantics: an inference rule that may look plausible is the *transitivity rule*

$$\frac{F \oplus G \qquad G \oplus H}{F \oplus H}.$$

However, it is unsound. Indeed, consider a signature with two distinct ground atoms p, q, and the following instance of the transitivity rule:

$$\frac{p \oplus p \vee q \qquad p \vee q \oplus q}{p \oplus q}.$$

The two premises are both entailed by the empty set, as they are derivable by applying the implication rule and symmetry rule to the tautologies

$$p \to (p \vee q),$$
$$q \to (p \vee q).$$

However, we have already shown that $p \oplus q$ is not entailed by the empty set, so the premises do not entail the conclusion.

Chapter 4

Establishing Coherence in DRT

The second contribution that we will outline is in bringing the Correlation Calculus closer to our current understanding of language. The translations of the sentences used for the examples in Bailey *et al.* [2015] were somewhat ad-hoc; instead we would like to extract first-order and correlation formulas from Winograd Schema sentences in a principled way. For that, we will make use of Discourse Representation Theory, which is a well-established theory of meaning representation for discourses—i.e., bodies of text with more than one clause.

So far we have discussed principled ways of

(i) constructing logical forms from discourses (Section 2.4), and

(ii) using correlation formulas to answer WSC questions (Section 2.3).

What remains is to fill the gap between the DRS for a Winograd Schema sentence and the correlation formula we use to find the answer. We will discuss a method to do this for certain kinds of sentences: in particular, those where two affirmative clauses are connected by "because." Then in Chapter 5 we will discuss how the approach may (or may not) generalize to other kinds of sentences.

4.1 DRSs for "Because" Sentences

The development of DRT in Kamp and Reyle [1993] has very little to say on words like "because," which connect clauses together. Part of this is because while "because" may have an interpretable truth-conditional meaning relating to *causation*,

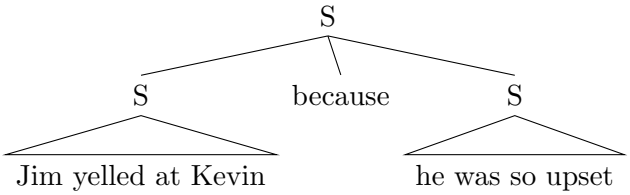
it is a member of a class of words—*but*, *although*, *however*, *so*, et cetera—whose meanings aren't so clear-cut.

These words, seen as indicating higher level *rhetorical* or *discourse relations* (rather than the *predicative* relations denoted by verbs), are treated more fully by the extension of DRT called *Segmented Discourse Representation Theory* (Asher and Lascarides [2003]). This system is much more complex, and involves a grammar of the whole discourse that connects clauses by rhetorical relations such as *Explanation*, *Elaboration*, and *Contrast*, among others. In SDRT, "because" indicates *Explanation*, which is then related semantically to a *causes* predicate. We will instead translate directly to a *because* predicate, and show how to extract a correlation formula from the resulting DRS. This approach is better suited to the Winograd Schema Challenge because it does not introduce unnecessary complexity[1] for the simple discourses in Winograd Schema sentences. But in light of the following analysis, it may turn out that the system *could* generalize to more complex discourses (see Chapter 5).

We will first illustrate how to handle "because" sentences by an example:

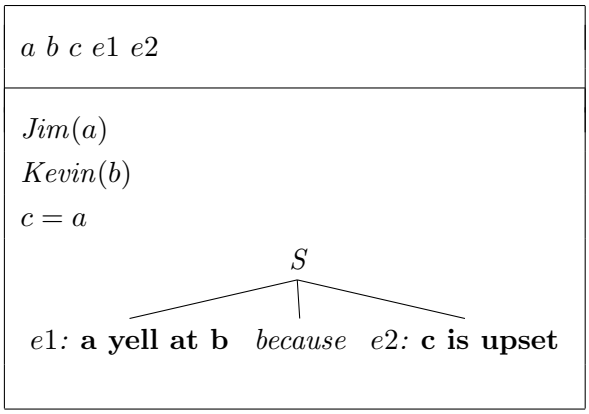
(15) Jim yelled at Kevin because he was so upset.

Whatever the details of the syntactic analysis, any reasonable constituency parse would label *Jim yelled at Kevin* and *he was so upset* as constituents. So we will assume a simplified analysis with the understanding that it could be easily extended to a more comprehensive account:



In constructing a DRS from this sentence, we may first ask what the original construction algorithm would give us. Since there is currently no rule for the word *because*, we would simply apply all of the possible rules to the clauses in the subordinate S nodes. The result looks as follows:

[1]While the SDRT system, in particular the problem of discourse parsing, is decidable, it may not be tractable to compute in practice.



(in which we have already resolved the anaphoric c, and ignored adverbs and tense). The last condition is *irreducible*—it represents an (annotated) tree structure to which no more rules can be applied.

Recall that DRT assumes a direct correspondence between the argument structure of a verb and the predicate invoked in its meaning representation (see Section 2.4.4). This is how irreducible conditions are normally handled, semantically speaking. There isn't such an obvious correspondence for these more complex conditions, but a translation is nonetheless apparent. To show it by example, the translation would result in the following:

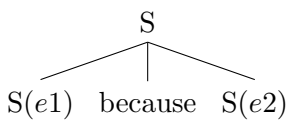
(16)

$a\ b\ c\ e1\ e2$
$Jim(a)$ $Kevin(b)$ $c = a$ *e1:* **a yell at b** *e2:* **c is upset** $because(e1, e2)$

This can be readily abstracted into a general rule.

(17) **Construction Rule CR.Because.**

Triggering Configuration.



Procedure.

(i) Add the three following DRS-conditions to the DRS:
- the S($e1$) subtree,
- the S($e2$) subtree, and
- the condition $\mathit{because}(e1, e2)$.

(ii) Delete the original tree containing the triggering configuration.

This rule works generally when "because" connects affirmative clauses. We will delay a discussion of the problems of this approach to Chapter 5, and for now move to how to extract the information we will use in the process of establishing discourse coherence from a DRS constructed by such a rule.

4.2 Extracting Correlation Formulas from DRSs

From a formula with a *because* condition, we will extract a correlation formula, and another formula we will call the *background information*. The process of extracting a correlation formula is fairly straightforward. Consider for example the DRS (16). We first locate the *because* condition: $\mathit{because}(e1, e2)$. We take this to indicate that the *positive correlation* relation should hold between the events denoted by its arguments $e1$ and $e2$. We then find these events in the DRS, $e1$: **a yell at b** and $e2$: **c is upset**. We have no need for the event-denoting discourse referents in the resulting correlation formula, since we are trying to express a correlation between the events denoted by these referents, rather than between statements *about* the particular referents. So for a condition $e : \gamma$, we use the translation $[[\gamma]]$ given in Section 2.4.3. Thus the correlation derived from this DRS will be

$$\mathit{yell\text{-}at}(a, b) \oplus \mathit{upset}(c). \tag{4.1}$$

The "dropping" (in some sense) of the event-denoting discourse referent facilitates the use of commonsense knowledge about general relationships instead of knowledge about particular events. The result looks very similar to the ad-hoc formulation in Bailey *et al.* [2015]. However, the origins of the predicates in use have become clear: they are event-denoting predicates with the event referent dropped. So they are either:

(i) full predicate-argument structures of verbs (e.g., yell at), or

(ii) full predicate-argument structures of adjectival descriptions that appear as the complement in a copular statement (e.g., upset, proud of).

However, this does leave modal operators unexplained: should they be represented explicitly in the FOL formulas, or should a modal logic be used underlying correlation formulas? There are formulations of DRT acounting for modality (see Geurts [1999]), but it changes the semantics and it is not clear what this would mean for the correlation calculus. We leave the issue unresolved in this work.

One other nuance needs to be clarified at this point: in formula (4.1), a, b, and c should not be free variables. If they are syntactically treated as variables, they will be free in the correlation formula, and background information from the DRS would either not apply to them or be incorrectly universally quantified. So in our translation to FOL, we will need to treat a, b, c, and all other discourse referents as object constants. This will be done explicitly by deleting the discourse referents in the universe of the main DRS in the procedure for the extraction of correlation formulas and background information. However, only the discourse referents from the *main* (i.e., outermost) DRS should be treated as object constants. To retain the proper meanings of all of the DRS-conditions, discourse referents introduced in subordinate DRSs will be treated as bound variables as is normal in the translation to FOL.

Finally, we will outline the procedure explicitly.

1. Let K be a proper DRS confined to the vocabulary V and the set R of discourse referents, such that K contains the following three DRS-conditions:

$$\mathit{because}(e1, e2) \qquad e1 : \gamma_1 \qquad e2 : \gamma_2,$$

   where $e1$ and $e2$ do not appear in γ_1, in γ_2, or anywhere else in the DRS, and γ_1 and γ_2 are instances of predicate-argument structures.

2. Let $\delta_1, \ldots, \delta_n$ be the DRS-conditions of K other than the *because* condition and the two event descriptions.

3. Then the correlation formula extracted from K will be

$$[[\gamma_1]] \oplus [[\gamma_2]]$$

and the background information will be

$$\bigwedge_{i=1}^{n} [[\delta_i]].$$

The two formulas will be understood as being over a signature containing the predicate symbols of V and object constants including the names in V and the discourse referents in the universe of K. Thus neither formula will have free variables.

So the DRS (16) yields the correlation formula (4.1):

$$\textit{yell-at}(a, b) \oplus \textit{upset}(c)$$

and background information

$$a = \textit{Jim} \wedge b = \textit{Kevin} \wedge c = a.$$

Establishing coherence One last unanswered question is how we will use DRSs and their correlation formulas to establish coherence and ultimately resolve anaphora. Here is the crucial point: An interpretation of a *because*-sentence is **coherent** if its corresponding correlation formula can be proven from its background information and commonsense knowledge.

The background information in the DRS corresponds to what Bailey *et al.* [2015] called *presuppositions*. In that case, the term was appropriate since the only information with that role in the derivations was originally the content of definite descriptions (which, in DRT, are considered presuppositional). But in the DRSs we constructed, we don't define an *a priori* way to distinguish between accommodated presuppositional content like that for proper names and not-at-issue propositional content like the contents of a relative clause. That is the reason we have adopted the broader term *background information*. But it is no coincidence that our approach ends up looking very much the same as the approach in Bailey *et al.* [2015], because their use of presuppositions was made with DRT in mind.

We view the solution of Winograd Schema Challenge questions as the byproduct of the process of establishing coherence. So to define our proposed approach to Winograd Schemas, we will first propose explicitly how to establish coherence in *because*-sentences.

This process will not only require a database of commonsense knowledge, but

also a system for doing automated theorem proving in the Correlation Calculus. This is yet another issue that we have left to future work; the exact computational properties of the correlation calculus are not yet established.

(18) **Procedure** (Establishing coherence in *because*-sentences).

Input.

- A non-negated statement consisting of two clauses connected by *because*, with a single ambiguous pronoun.
- A knowledge base of first-order sentences and correlation formulas.
- An automated prover for the Correlation Calculus.

Output. A set of DRSs, deemed coherent interpretations of the discourse.

Steps.

1. Construct the possible DRSs for the statement, using the system outlined in Section 2.4.4 augmented with the new *because* rule (item 17).
2. For each DRS, use the automated prover to attempt to prove its correlation formula from axioms in the knowledge base and its background information.[2] If the proof is successful, the DRS is deemed a coherent interpretation.[3]
3. Return all DRSs deemed coherent interpretations.

We may now outline a way to use the set of coherent interpretations to answer Winograd Schema questions of the *because* form:

1. First, apply procedure (18) to yield a set of coherent interpretations (DRSs) of the discourse. In the process, take note of the discourse referent introduced in the **CR.PRO** rule that was applied to the anaphor. Call it z.

[2] We leave open the question of how to implement the failure of a proof; a sensible option might be limiting the prover by an empirically-determined timeout.

[3] There is one more step we might add in principle at this point. Recall that a correlation formula $F \oplus G$ is understood to mean that F and G are positively correlated *or the truth or falsity of F or G is known to the hearer.* In the latter case, it may be reasonable to say the correlation formula is no longer evidence of correlation, and thus is not fit to establish discourse coherence. How to determine coherence in that case, however, is not clear. But the convenient fact is that this seems like it will rarely be the case. Since the formulas on either side will include new object constants, relevant knowledge about them would have to be deducible only from universally quantified axioms and the background information. It seems that this will rarely be enough information to cause problems, and will never be an issue in the tiny discourses of the Winograd Schema Challenge.

2. Examine the two answers to the Winograd Schema question. They will take the form of definite noun phrases. For each answer A, let the formula F_A be defined as follows:

   - If A is a proper name π, F_A is

     $$z = \pi.$$

   - If A is a definite noun phrase of the form *the* N, where N is a common noun, F_A is

     $$N(z).$$

   Other cases, such as more complex phrases including adjectives or relative clauses, may be handled similarly if we take a look at their treatment in the DRS Construction Algorithm.

3. For each coherent DRS K, ask the following question for each A:

   Does the formula F_A follow from the background information of K?

   (This is essentially the question "does it follow that *it* is A?") Attempt to answer this with an automated prover for the Correlation Calculus.

4. If for all coherent DRSs, one proof succeeded and the other did not, and if this proof corresponded to the same answer in all cases, then choose that answer to the Winograd Schema Challenge question.

4.3 Examples of Coherence

In this section, we will give justifications for three Winograd Schema answers from the list compiled by Ernest Davis. For these examples, we will not exactly follow the procedure outlined in the previous section. Instead of constructing *all* possible DRSs, we will only prove the coherence of the DRS that happens to be the correct interpretation. Nevertheless, it is clear in each case that with the background information and commonsense axioms provided, the coherence of the incorrect answer cannot be proven. Unfortunately, in order to establish that concretely, we would need to establish broad claims about what should *n*ot be entailed by the knowledge base. This will be discussed briefly for each example.

Yelling at Kevin Schema 18 on Davis's list reads:

(19) Jim [yelled at/comforted] Kevin because he was so upset. Who was so upset?

We will justify the answer *Jim* for the first instance as follows. First, we construct the DRS (for the correct interpretation):

$a\ b\ c\ e1\ e2$
$Jim(a)$
$Kevin(b)$
$c = a$
$e1$*:* **a yell at b**
$e2$*:* **c is upset**
$because(e1, e2)$

.

Then we extract the correlation formula:

$$\textit{yell-at}(a, b) \oplus \textit{upset}(c).$$

We then list the background information

1. $a = Jim$
2. $b = Kevin$
3. $c = a$

and the relevant commonsense knowledge:

4. $\textit{yell-at}(x, y) \oplus \textit{upset}(x)$.

The derivation continues as follows:

5. $\textit{yell-at}(a, b) \oplus \textit{upset}(a)$.	by substitution from 4
6. $\textit{upset}(a) \leftrightarrow \textit{upset}(c)$	entailed by 3
7. $\textit{yell-at}(a, b) \oplus \textit{upset}(c)$	by replacement from 5 and 6.

Here we required the commonsense knowledge that *x yelling at y is positively correlated with x being upset.* The opposite answer would be provable if and only if the

formula

$$\textit{yell-at}(a, b) \oplus \textit{upset}(b)$$

were provable from the knowledge base and background knowledge. This is actually quite a reasonable axiom—which raises a problem. See Section 5.3 for details.

Envying Martin Schema 20 on Davis's list reads:

(20) Pete envies Martin [because/although] he is very successful. Who is very successful?

We will justify the answer *Martin* for the first instance. First, we construct the DRS:

$a\ b\ c\ e1\ e2$
$\textit{Pete}(a)$ $\textit{Martin}(b)$ $c = b$ $e1$*:* **a envy b** $e2$*:* **c is successful** $\textit{because}(e1, e2)$

.

Then we extract the correlation formula:

$$\textit{envy}(a, b) \oplus \textit{successful}(c).$$

We then list the background information

1. $a = \textit{Pete}$
2. $b = \textit{Martin}$
3. $c = b$

and the relevant commonsense knowledge:

4. $\textit{envy}(x, y) \oplus \textit{successful}(y)$.

The derivation continues as follows:

5. $envy(a, b) \oplus successful(b)$	by substitution from 4
6. $successful(b) \leftrightarrow successful(c)$	entailed by 3
7. $envy(a, b) \oplus successful(c)$	by replacement from 5 and 6.

Here we required the commonsense knowledge that *x envying y is positively correlated with y being successful.* The incorrect answer *Pete* would be provable if

$$envy(a, b) \oplus successful(a) \tag{4.2}$$

were provable from the knowledge base. One way to ensure that this is not the case is to ensure that

$$envy(a, b) \ominus successful(a)$$

is entailed by the knowledge base—this axiom seems reasonable, expressing that successful people tend to be less envious. Given that the formulas on either side are not entailed (which they definitely should not be, given that a and b are new object constants) this formula is contradictory (by the probabilistic semantics) to the correlation formula corresponding to the answer we would wish to avoid. Thus, because the calculus is sound, as long as the knowledge base is consistent, the undesirable formula (4.2) would not be derivable.

Ceding the Election Schema 139 on Davis's list reads:

(21) Kirilov ceded the election to Shatov because he was [more/less] popular. Who was [more/less] popular?

We will justify the answer *Shatov* for the first instance of this schema. First, we construct the DRS:

$e1\ a\ b\ c\ e2\ d\ f$
$Kirilov(a)$ $Shatov(b)$ $election(c)$ $d = b$ $f = a$ $e1$*:* **a cede c to b** $e2$*:* **d is more popular than f** $because(e1, e2)$

.

Here, we have assumed a slightly more complicated DRS that relies on the more complicated syntactic analysis that would arise from the comparative *more popular*. This is comparing *against* something, and whether the syntactic analysis resolves the issue as one of anaphora or of ellipsis, the result must involve a referent to denote that thing. We use f. Then this problem involves resolving what amounts to *two* anaphors: d and f. Of all of the possible choices, we will jointly justify the choice of $d = b$ and $f = a$.

So the next step is to extract the correlation formula:

$$cede\text{-}to(a, c, b) \oplus more\text{-}popular(d, f).$$

We then list the background information

1. $a = Kirilov$
2. $b = Shatov$
3. $election(c)$
4. $d = b$
5. $f = a$

and the relevant commonsense knowledge:

6. $cede\text{-}to(x, y, z) \wedge election(y) \oplus more\text{-}popular(z, x)$.

The derivation continues as follows:

7.	$\textit{cede-to}(a, c, b) \wedge \textit{election}(c) \oplus \textit{more-popular}(b, a)$	by substitution from 6
8.	$(\textit{cede-to}(a, c, b) \wedge \textit{election}(c)) \leftrightarrow \textit{cede-to}(a, c, b)$	entailed by 3
9.	$\textit{cede-to}(a, c, b) \oplus \textit{more-popular}(b, a)$	by replacement from 7 and 8
10.	$\textit{more-popular}(b, a) \leftrightarrow \textit{more-popular}(d, f)$	entailed by 4 and 5
11.	$\textit{cede-to}(a, c, b) \oplus \textit{more-popular}(d, f)$	by replacement from 9 and 10.

Here we required the commonsense knowledge that *x ceding an election z to y is positively correlated with y being more popular than z.* The knowledge that would prevent the incorrect answer might as well be another axiom:

$$\textit{cede-to}(x, y, z) \wedge \textit{election}(y) \ominus \textit{more-popular}(x, z).$$

This in turn may even be justified by the simple facts

$$\forall x \forall y (\textit{more-popular}(x, y) \leftrightarrow \textit{less-popular}(y, x))$$

and

$$\forall x \forall y (\textit{more-popular}(x, y) \leftrightarrow \neg \textit{less-popular}(x, y))$$

(though the latter admittedly ignores the case that the two are equally popular).

In this example, the Correlation Calculus has solved a slightly more complex decision than the WSC problem poses: it has resolved two anaphors instead of just the one. This shows how the approach may have explanatory power for discourse coherence more generally, rather than just in the Winograd Schema Challenge.

Nevertheless, the formulation of this last example (indeed, all of these examples) might raise some questions. For instance, is it correct to include the *election* constraint on the *z* argument to *cede-to*? Or, should we instead say that *cede* is polysemic, and has one specific meaning (and thus, predicate in the first-order signature) corresponding to ceding *elections* in particular (thus avoiding the need for the *election* constraint altogether)? Or perhaps worse, the *more-popular* predicate seems to miss out on the deeper common structure of gradable adjectives or comparatives. These are important questions for the knowledge-representation side of a system to solve the Winograd Schema Challenge. But they are not as important to the Correlation Calculus: the fact that these proofs *can be done* in a relatively

simple system is a *first step* to showing that they may be doable in systems with richer ontologies. We will discuss some of these issues in Chapter 5.

Chapter 5

Discussion

5.1 Beyond "Because"

Though we developed the system with the word "because" in mind, not all of the Winograd Schema problems take this form. However, there's a fairly straightforward way of generalizing the approach. In the first 100 Winograd Schemas of the collection compiled by Ernest Davis,[1]

- 96 put the pronoun in a separate independent clause,
- 71.5 of the answers seem to be justifiable by positive correlation,[2] and
- of those justifiable by positive correlation, the clause containing the pronoun is connected to the discourse by one of the following: a full stop, a semicolon, *because, so, if, since, when, but, now, and, and then, until.*

So it seems that generalizing the approach simply to apply to more connectives could be a fruitful endeavor. The task would not be simple: for example *but*, which appears above, might be expected to indicate negative correlation. But examples with *but*, 4.5 seemed justifiable by negative correlation, 1 seemed justifiable by positive correlation, and 6 did not seem to be related to correlation—so it is clearly a difficult case. The connectives *though* and *although*, though uncommon, seemed to indicate negative correlation in their 2.5 occurrences. All of the other connectives mentioned above appeared in no schemas that could be justified by negative correlation, and

[1] `https://www.cs.nyu.edu/davise/papers/WS.html`

[2] Since we are counting each schema once, a single schema instance counts as 0.5.

mostly schemas that could be justified by positive correlation. For details, the annotations may be found at `cs.utexas.edu/~julianjm/wsc-annotations.txt`.

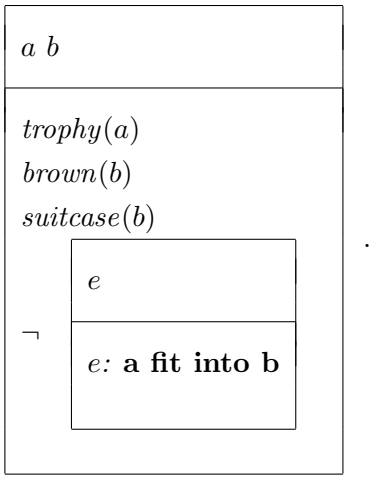
Even in cases where positive correlation seems to apply, the technique discussed in this thesis be difficult to generalize to cases where the observed correlation is not between statements of eventuality, or the correlated statements appear far apart in the discourse. But the most difficult test of the ability of this method to generalize will be how it may be adapted to deal with richer and more general systems of commonsense knowledge, or the kind of commonsense knowledge that may be obtained from data. What these systems or this knowledge may look like is an open question for future work.

However, the Correlation Calculus is not just a system for justifying WSC answers. It is primarily a system for establishing discourse coherence on the basis of commonsense knowledge. Especially because of the state of the rest of the technology needed to actually apply the approach to the WSC, it is more useful to think of the Correlation Calculus approach as a contribution to our thinking on discourse coherence.

5.2 Negation

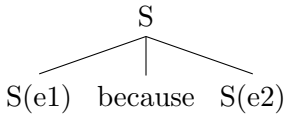
One issue that we have conveniently skirted is negation. This is especially important for the Winograd Schema Challenge, since almost any WSC problem could be phrased (and many of them *are* phrased) with negation. The problem with negation arises very early in the process: incorporating the *because* predicate in a DRS.

The problem is that our use of the *because* predicate requires an event-denoting discourse referent for each of the two clauses being related. However, take a look at the DRS for *the trophy doesn't fit into the brown suitcase*:



There is no accessible discourse referent representing the event denoted by the statement. Rather, the DRS-condition with the negation signifies that no such referent may exist.

Indeed, in a complex sentence with *because* and negation, our construction rule **CR.Because** is not applicable. Its triggering configuration looks like



where $e1$ and $e2$ are event-denoting discourse referents. This configuration, in the system developed by Kamp and Reyle [1993], will only occur when both clauses are affirmative. In the presence of negation, the annotation on an S-node will instead be a *time*-denoting discourse referent, denoting the interval of time during which no event described by the sentence occurred. This leaves us stuck. It's not entirely clear how to proceed, either. Indeed, it seems that a condition like *because*, which relates two statements to each other, simply can't be expressed about negative statements in this version of DRT.

A couple of alternatives arise at this point. First, there is an extension of DRT, developed by de Swart and Molendijk [1999], which assigns state-denoting discourse referents to negated statements. Adopting their system may make it possible to recover correlation formulas. Second, we may decide to change from using *event*-denoting referents as the anchor for correlation formulas to using *time*-denoting referents. This may complicate the process of extracting a correlation formula, but it might also have the advantage of allowing us to encode temporal information in our correlation formulas and commonsense knowledge (see Section 5.3).

We leave the exploration of these directions to future work.

5.3 Correlation, Causation, and Time

Solving a Winograd Schema question with the Correlation Calculus involves more than just establishing the coherence of the correct interpretation; it requires establishing that the other interpretations are not coherent.

Therefore in our analysis of Winograd Schema questions, we need to show not only that the solutions can be supported by reasonable commonsense knowledge, but also that the *incorrect* choice should *not* be supportable from reasonable commonsense knowledge. This presents a problem with, for example, the *yelling at Kevin* schema instance:

> Jim yelled at Kevin because he was so upset.

We make use of the axiom

$$\textit{yell-at}(x, y) \oplus \textit{upset}(x)$$

to prove that the answer is *Jim*, but the axiom

$$\textit{yell-at}(x, y) \oplus \textit{upset}(y)$$

may seem just as reasonable: *x yelling at y* may be positively correlated with *y being upset*.

But this axiom would license the choice of *Kevin* as the answer in this problem. Thus perhaps the correlation calculus *isn't* fit to solve this basic WSC question. The factor that's missing is exactly what we lose when we translate *because* to positive correlation: the directionality of causation. What's really going on is that $\textit{upset}(x)$ causes $\textit{yell-at}(x, y)$, and $\textit{yell-at}(x, y)$ causes $\textit{upset}(y)$. Both manifest as a positive correlation, but because the problem says "because he was so upset" only the first is relevant.

This would seem to be a fundamental limitation of correlation—which, as any scientist knows, does not imply causation—but not all is lost. There may be a way to implicitly encode causality in correlation formulas, and it makes use of the relationship between causality and time.

Suppose instead of the formula *yell-at*(x, y) to represent x yelling at y, we used *yell-at*(x, y, t) to represent x yelling at y *at time* t. Then, the relevant information for this problem could be encoded as follows:

$$\textit{yell-at}(x, y, t_2) \wedge (t_1 < t_2) \oplus \textit{upset}(x, t_1) \wedge (t_1 < t_2).$$

The knowledge that yelling at someone is positively correlated with *making* them upset could be encoded by using the opposite order of the time-denoting discourse referents. Finally, the connective "because" could involve a constraint that the cause cannot come only after the result.

Three issues are worthy of noting with this approach. First, that there already exists a rich temporal ontology in DRT, and such an approach may benefit from making use of it—for example, we might not want to say so much that x being upset *precedes* x yelling at y, but that x yelling at y is temporally *contained* within x being upset. DRT's temporal ontology lets us express this.

Second, in many cases (including, arguably, the yelling example) there is ambiguity about the temporal structure of the discourse. Our process of establishing coherence, which relies on testing various antecedents of a pronoun, might possibly be extended to also test possible *temporal structures*. This would mean the Correlation Calculus might be fit for more complex discourse interpretation tasks, of the kind that Segmented DRT was designed to address.

Third, the interaction of positive versus negative correlation with the relative temporal structure of events may lead to a richer typology of rhetorical relations that could relate in a meaningful way to the relations in the literature.

All of these are issues for future work.

5.4 Conclusion

In this thesis, we have presented several developments of the Correlation Calculus, proposed by Bailey *et al.* [2015] as an approach to establishing discourse coherence on the basis of commonsense knowledge. First, we have given it a precise semantics, which is useful in establishing some of the properties of the calculus. Second, we have systematically integrated the use of the Correlation Calculus with Discourse Representation Theory to explain a general method of establishing coherence for some very specific discourse structures. This has filled some of the conceptual holes

with the approach of the Correlation Calculus to the Winograd Schema Challenge.

The main contribution of the calculus is to the theory of discourse coherence. There are not many competing theories: they range from an implemented system based purely on commonsense knowledge (Inoue *et al.* [2012]) to a system that is highly sensitive to a nuanced, complex model of discourse structure (Asher and Lascarides [2003]). Both of the approaches have merits, and we believe we have taken the first steps in establishing a middle-of-the-road approach that has the following favorable properties:

- The method is sensitive to discourse structure,
- the calculus has a simple and comprehensible monotonic semantics, and
- the problem of knowledge representation is well-defined.

Whether the approach would be able to be applied to the WSC in practice is an open question. But we believe the Correlation Calculus is in the spirit of the Challenge, and the *science of AI* as regarded by Levesque [2014]. That is, it is a step towards *explaining the phenomena* of human behavior on Winograd Schemas.

Bibliography

Nicholas Asher and Alex Lascarides. Lexical disambiguation in a discourse context. *Journal of semantics*, 12(1):69–108, 1995.

Nicolas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

Daniel Bailey, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. The Winograd Schema Challenge and reasoning about correlation. In *Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*. AAAI Press, 2015. `http://www.cs.utexas.edu/users/ai-lab/?wsc15`.

Patrick Blackburn and Johan Bos. Representation and inference for natural language: A first course in computational semantics, 1999.

Johan Bos. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications, 2008.

Tejas Ulhas Budukh. An intelligent co-reference resolver for Winograd schema sentences containing resolved semantic entities. Master's thesis, Arizona State University, 2013.

Robin Cooper. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112, 2005.

Donald Davidson. The logical form of action sentences. In *The Logic of Decision and Action*, pages 81–120. University of Pittsburgh Press, 1967.

Henriette de Swart and Arie Molendijk. Negation and the temporal structure of narrative discourse. *Journal of Semantics*, 16(1):1–42, 1999.

Gerald Gazdar, E. Klein, G. Pullum, and I. Sag. *Generalized Phrase Structure Grammar*. Blackwell, Oxford, 1985.

Bart Geurts. *Presuppositions and Pronouns*. Elsevier, 1999.

I. Heim and A. Kratzer. *Semantics in Generative Grammar*. Blackwell Textbooks in Linguistics. Wiley, 1998.

J Hobbs, M. Stickel, D. Appelt, and P. Martin. Interpretation as abduction. *Artificial Intelligence*, 63:69–142, 1993.

Jerry Hobbs. Coherence and coreference. *Cognitive Science*, 3:67–90, 1979.

Jerry R. Hobbs. Ontological promiscuity. In *Proceedings of 23rd Annual Meeting of the Association for Computational Linguistics - ACL 85*, pages 61–69, Chicago, Illinois, USA, 1985. University of Chicago.

Naoya Inoue, Ekaterina Ovchinnikova, Kentaro Inui, and Jerry Hobbs. Coreference resolution with ilp-based weighted abduction. In *COLING*, pages 1291–1308, 2012.

Hans Kamp and Uwe Reyle. *From discourse to logic*, volume 1,2. Kluwer, 1993.

Hans Kamp and Uwe Reyle. A calculus for first order discourse representation structures. *Journal of Logic, Language and Information*, 5(3/4):297–348, 1996.

Makoto Kanazawa. Weak vs. strong readings of donkey sentences and monotonicity inference in a dynamic setting. *Linguistics and Philosophy*, 17(2):109–158, 1994.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman. Coherence and coreference revisited. *Journal of Semantics*, 2008.

Alex Lascarides and Nicholas Asher. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493, 1993.

Alex Lascarides and Nicholas Asher. Segmented Discourse Representation Theory: Dynamic semantics with discourse structure. In *Computing Meaning*, pages 87–124. Springer, 2007.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the

conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Comput. Linguist.*, 39(4):885–916, December 2013.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. In *Proceedings of International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 2012.

Hector J Levesque. On our best behaviour. *Artificial Intelligence*, 212:27–35, 2014.

Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, 2012.

Aarne Ranta. *Type-Theoretical Grammar*. Oxford University Press, 1995.

Arpit Sharma, Nguyen H. Vo, Shruti Gaur, and Chitta Baral. An approach to solve Winograd Schema Challenge using automatically extracted commonsense knowledge. In *Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*. AAAI Press, 2015.

Arpit Sharma. Solving Winograd Schema Challenge: Using semantic parsing, automatic knowledge acquisition and logical reasoning. Master's thesis, Arizona State University, 2014.

Terry Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1 – 191, 1972.