Attitudes and Changing Contexts

Robert van Rooij

Contents

Preface

This book centres on three themes: first, the *pragmatics* of natural language interpretation; second, the *externalist* view of content; and third, the analysis of *counterfactual dependencies* and *robustness*. These three themes will come back again and again in the various chapters; they are relevant to the analysis of attitude attributions, anaphoric and presuppositional dependencies, and conditional sentences.

The first theme is related to the *pragmatics* of natural language, or better the *semantic-pragmatic interface*. Pragmatics studies the two-way interaction between *utterances* and the *contexts* in which they are made. The relation is one of interaction, not only because *what* is expressed by utterances depends on context, but also because utterances *change* the context of evaluation.

It is obvious that *what* is expressed by an utterance in which an indexical expression or pronoun occurs depends on context. It follows that to account for this context dependence of utterances, contexts have to contain enough information to determine the referents of referentially-used expressions. One of the major claims of this book will be that contexts also influence what is expressed or communicated by sentences that do not contain such obvious context-dependent expressions. As I will stress, this will be the case in particular for *attitude attributions* and utterances of *conditional sentences*. What is communicated by such sentences depends on what is believed and presupposed by the participants in a conversation. This latter kind of context dependence will obviously put additional constraints on how contexts should be represented.

Because a context should represent what the participants in a conversation believe and presuppose, an utterance will also influence the context of evaluation. That is, the *context changes* after an utterance has been made. This context change can happen in a *direct* way. For instance, when a speaker has made an assertion and nobody protests, it can be assumed that the content of what was asserted is now accepted by the participants in a conversation, and can now be presupposed during the rest of the discourse. But utterances can also influence contexts in more *indirect* ways, by means of *accommodation*. It is normally assumed that a rational speaker can only *appropriately* make certain utterances when certain conditions are fulfilled. For instance, a speaker can normally make an assertion appropriately only when he himself (i) believes the content of the assertion, and (ii) assumes that the content of his assertion is not yet commonly assumed, i.e. presupposed. It need not always be clear to the hearer what the speaker presupposes and believes.

However, when the speaker makes an assertion that normally can be made appropriately only by a rational speaker when he has certain beliefs and presuppositions, the hearer can conclude that the speaker indeed had these beliefs and presuppositions, and *accommodates* the context accordingly.

We have seen that contexts are used to determine both *what* is expressed by a context-dependent utterance and whether an utterance is made *appropriately* or not. The central idea behind any appropriate pragmatic theory, such as the recently developed theories of context change to be discussed in chapter 2, is that there is a *single* notion of context that contains enough information about the conversational situation to determine both the content and the appropriateness of utterances; and that both kinds of information modelled by this single context change during a conversation in an interactive way.

What kind of information determines the reference of referentially-used expressions? The *second theme* of this book is related to the claim that the meaning and content of linguistic expressions should be explained in terms of the intentions, beliefs, and conventions of language users; and that the content of what is intended, believed, and presupposed is to be partially explained in *externalistic* terms. If we assume that the content of our attitudes should be explained in externalistic terms, we can make our claims compatible with a causal theory of reference for which convincing arguments have been given by Kripke and others. That is, in this way we can make this causal theory of reference compatible with the intuition that reference should be explained in terms of what speakers do by their use of a term, and not by properties of the term itself. I will defend this causal theory of reference not only for proper names and common nouns, but also for certain uses of definite descriptions and anaphoric pronouns.

Aboutness is sometimes explained in terms of *actual* causal relations between the information states and the objects they are about. I will argue, however, that the causal or information-theoretic analysis of aboutness should in general be cashed out in terms of *counterfactual relations* between the information states and the object or information that this state is about. How to analyze such counterfactual relations, and how to use these analyses to account for certain attitude attributions, is the *third main theme* of this book.

The three themes described above will show up again and again in the different chapters of this book. Chapter 1, entitled 'Belief and Belief Attribution', defends the externalist theory of content and belief. One of the consequences of the causal/externalistic account of content is that the *semantic values* of proper names and indexicals are equated with their references. But as is well known, this property of these externalist theories gives rise to a number of problems, especially when epistemic contexts are involved. A significant part of chapter 1 is devoted to motivating and explaining a three-part solution to these problems. The first part of the solution will be to take seriously the idea that although the content of a mental or linguistic representation depends on external conditions, it might be unclear for believers, or for participants in a conversation, what the relevant external conditions are. This part of the solution will make use of the Stalnakerian technique

of *diagonalisation* in a *two-dimensional* framework. The second part of the solution will make use of a *counterpart theory* to account for Quine's *double vision* problem, where Ralph believes of Ortcutt that he is a spy, and believes of Ortcutt that he is not a spy, although intuitively he need not be *internally* inconsistent. The third and perhaps most important part of the solution will account for the intuition that belief attributions are extremely *context dependent*: what is expressed/communicated by a belief attribution depends not only on (i) the referent of the indexical or pronoun used in the embedded clause, but also on (ii) the representative of the individual the attributed belief is about, and (iii) the set of relevant possibilities in terms of which the agent's belief state is defined. All three depend crucially on context.

In chapter 1, I seek to reconstruct the Stalnakerian position with respect to content and belief attributions. It is partly built on the insight – due to Kaplan, Stalnaker and others – that it is good to make a conceptual distinction between two kinds of facts: (i) facts about the subject matter of conversation, and (ii) facts about linguistic and speech conventions and the conversational situation itself. This conceptual distinction will be used extensively in the following three chapters about anaphora and presuppositions.

In chapter 2, I will account for anaphoric relations across sentential boundaries on the basis of the intuition that pronouns are normally used referentially and the assumption motivated in chapter 1 that referring is something done by speakers with their use of a term and not by the term itself: which object is referred to depends on the intention of the speaker. Kripke (1977) taught us that a distinction must be made between *general* and *specific* intentions. I will argue that for pronouns it is normally the specific intention that counts. Speakers normally refer back with their use of a pronoun, or short description, to the *speaker's referent* associated with the indefinite that figures as its syntactic antecedent. In this chapter I will show that by means of diagonalisation such an analysis can be pushed further than many have supposed; and that in fact this analysis is close to, but not identical with, modern theories like Discourse Representation Theory (Kamp, 1981), File Change Semantics (Heim, 1982), and the more recent Dynamic Semantics due mostly to Groenendijk & Stokhof (1991). The reason is that participants in a conversation are normally not only unclear about the facts relating to the subject matter of conversation, but also about certain facts relating to the conversational situation itself.

Of course, sometimes a singular pronoun that takes an indefinite as its syntactic antecedent can be used appropriately although it does not refer to the specific speaker's referent of the indefinite. Sometimes it is only the general intention that counts. I will argue that to account for many of these cases we need *descriptive pronouns* in addition to referential pronouns. The former are pronouns that go proxy for a description recoverable from the sentence in which its syntactic antecedent occurs. In chapter 2 I will be concerned mainly with motivating this division of labour, implementing this analysis of pronouns in a dynamic theory of meaning, and using this two-tiered approach to account for phenomena problematic for the above-mentioned popular theories of anaphora.

In chapter 3, I discuss anaphoric relations across belief attributions, concentrating mainly on the problem of *intentional identity* made famous by Geach's Hob-Nob sentences. I will discuss how much of the popular view, which takes so-called unbound pronouns either as abbreviations for the antecedent clause or as variables bound by a dynamic existential quantifier, can be maintained. I will suggest, in fact, that this view cannot be maintained, and that Hob-Nob sentences give us an additional argument to take the notion of *speaker's reference* seriously in *semantics*.

In chapter 4, I assume that *presupposition* is a propositional attitude and that what is presupposed is what the speaker takes to be presumed common knowledge between speaker and hearer. As a result, presupposition should be given a *pragmatic* analysis: what a *sentence* presupposes should be explained in terms of what *speakers* normally presuppose by their use of sentences. According to the satisfaction approach to presupposition, every sentence should be interpreted with respect to a context, and this context should already contain the information that is presupposed by the interpreted sentence. Recently, this satisfaction approach to presupposition has been implemented within dynamic semantics, but it is well known that this straightforward implementation gives rise to certain empirical problems. In chapter 4 I make the satisfaction approach more compatible with the relevant data by (i) taking more seriously the idea of treating presupposition as a *propositional attitude*; (ii) assuming that there might be more information states around in terms of which a presupposition might be satisfied; and (iii) making use of *modal subordination*.

According to the Lewis/Stalnaker analysis of conditionals, the truth conditions of a conditional sentence depend crucially on the speaker's intentions. The speaker's intentions, together with the antecedent and other facts about the actual world, select the relevant world(s) with respect to which the truth value of the consequent, and thus of the whole conditional, is evaluated. Stalnaker tried to make a stronger claim: the formal properties of the function that does this selection should be explained in terms of the beliefs and presuppositions of language users. He proposed that the analysis of conditionals should be related to the analysis of belief revision. I will discuss this project in chapter 5 and give some attention to Lewis' triviality result, which showed that what is expressed by a conditional must be even more context-dependent than the original Lewis/Stalnaker analysis suggested, if conditionals are to be explained in terms of conditional beliefs. I will argue that most conditionals express propositions, but that the proposition expressed by an indicative conditional depends more directly on what is believed and presupposed by the speaker than the proposition expressed by a subjunctive conditional. Some examples will be discussed that show the context-dependence of conditional sentences, and which are traditionally thought of as being problematic for the Lewis/Stalnaker analysis. I will show that by making use of *diagonalisation* and *context change* these problematic examples can be accounted for appropriately.

In the final chapter, I make use of the analyses of conditionals, belief revision, and rational decision discussed in the previous chapter in order to account for the meaning

of some attitude verbs other than *believe*. To account for belief change we need a richer representation of belief states than is commonly assumed, and I will argue in chapter 6 that this richer representation is important for the analysis of, among others, evidential and buletic attitude attributions. A number of *evidential* verbs will be analyzed, for instance, in terms of *robustness* under belief revision. I will argue that this richer representation of belief states will also be useful for the analysis of attitudes of *desire*, in particular for the analysis of 'intention'.

Chapter 1

Content, belief and belief attributions

1.1 Introduction

According to the most straightforward account of belief attributions, the meaning of *a believes that A* in a particular context *c* is compositionally determined from the meaning of its parts in *c*. If it is assumed that meanings are assigned primarily to expressions, on this approach it seems that Frege's well known substitution puzzles, like those that will be presented in section 1.3, forces one to assume that meanings and contents are really very fine-grained entities, and that a belief state should be modelled in a very fine-grained way. The problem with this approach is that it seems hard to give any *independent* motivation for such a fine-grained notion of content. The alternative strategy would be to start out by giving a philosophically motivated notion of content, independent of belief attributions. Such an independent notion of content will then typically be a rather coarse-grained notion. The fact that so many belief attributions still seem to be true and appropriate is then explained, according to this alternative strategy, partly in terms of the intentions and presuppositions of the agent who is making the belief attribution. It is this latter strategy that I will be defending in this chapter.

The above mentioned philosophically motivated notion of content, I will argue, will be a combination of the pragmatic account of intentionality defended by, for instance, Ramsey (1931) and the causal information-theoretic account as proposed by, among others, Stampe (1977) and Dretske (1981). Both accounts motivate a rather coarse-grained analysis of belief states and of the content expressed by a sentence. According to both accounts, what the content of a belief state is depends on certain dispositional and counterfactual relations between intentional states and the world. In expressing such relations, statements that are truth-conditionally equivalent can be substituted for each other, which suggests that the possible world analysis of the content of belief states is the correct one. On the basis of this I will argue with Stalnaker that if we forget about the dynamics of belief, both the contents of belief states of agents, and the contents expressed by sentences can, and should, to a large extent, be modelled by sets of possible worlds.

The main part of this chapter will address the question of the extent to which the

causal and the pragmatic accounts of intentionality are compatible with each other. The causal account sometimes seems to predict a too specific and sometimes a too unspecific notion of content and object of belief. I will discuss these problems mainly by looking at the traditional questions of how to handle *de dicto*, *de se* and *de re* belief attributions within possible world semantics. I will argue that most problems can be accounted for when we (i) assume that the meanings of expressions are context dependent; (ii) separate questions about *attitude attribution* from questions about the *contents* of the attitudes themselves; and (iii) distinguish between an *object* a belief is about, and the body of *information* the agent has about this object.

In this chapter I start with a sketch of the possible world framework, with the notion of *proposition* defined in terms of it, and claim that belief states, and embedded sentences of belief attributions, should be modelled by such propositions. Before I motivate this analysis of belief and belief attributions, and defend it to obvious criticism, I first make a digression to the theory of reference. After stating the description theory of speaker's reference in section 3, I discuss the persuasive arguments Kripke and others have raised against this theory of meaning, and say something about their alternative causal theory of content. After sketching the pragmatic account of intentionality, I will show that similar arguments used against the description theory of meaning also indicate that the pragmatic account of intentionality has to be supplemented by a causal, or information-theoretic, account. I argue that both the pragmatic and the causal information-theoretic accounts of content indicate that belief states looked at from the agent's point of view should be individuated by truth conditions, and thus should be represented by sets of possible worlds. Then I discuss some well-known problems raised by the assumption that causality plays such an important role in mental and linguistic representations.

The remainder of this chapter is devoted to motivating and explaining a three-part solution to these problems. First, I argue that although the content of a mental or linguistic representation depends on external conditions, it might be unclear for believers, or for participants in a conversation, what the relevant external conditions are. Formally, I will argue that the meaning of an expression can be both index- and context-dependent, and that in a counterfactual reference-context referential expressions might have a different referent than in the actual reference-context. I will argue that with the Stalnakerian technique of *diagonalisation* some problems concerning beliefs and *de dicto* belief attributions can be solved. I give special attention to self-locating beliefs, because they show the impossibility of a purely descriptive account of content, and thus are, I believe, the greatest threat to a purely possible-world account of belief.

Unfortunately, diagonalisation cannot solve all problematic belief attributions. It can, for instance, not solve those problems for a formal theory that wants to take seriously the issues that Quine and others have raised for *de re* belief attributions where individuals are used to characterize a belief state. To account for this problem, I will argue that we need some kind of *counterpart theory* that allows for the possibility that, for instance,

one individual in one world has two distinct representatives in another, and that this is compatible with an account of content that is not purely descriptive. Such a counterpart theory is the second part of the strategy I want to defend.

But perhaps the most essential part of the strategy is an account of the extremely *context-dependent* nature of belief attributions. First, we need to account for the intuition that in different conversational situations, the same belief attribution can communicate different propositions, although the agent himself has not changed his mind. Second, I will argue that not only what is expressed by a belief attribution depends on the intentions and presuppositions of the attributer, but also how we should represent what the agent believes.

In Appendix A I formulate a double indexing counterpart semantics for modal logic, where I account in a formal way for most of the ideas argued for in this chapter.

1.2 Possible world semantics

Perhaps the main goal of semantics is to determine the *truth conditions* of statements, and to explain how these truth conditions are functionally dependent on the meanings, or semantic values, of their (direct) parts. The reason is that it seems reasonable to say that to know the meaning of a (declarative) sentence is to know what conditions have to be fulfilled to make the sentence true. To state these conditions in non-linguistic terms, this means that knowing the meaning of a sentence is to know under which *circumstances* it is true, where these circumstances can be thought of as the ways the world might have been, *possible worlds*. Thus, to know the meaning of a sentence is to be able to distinguish worlds where the sentence is true from those where it is false, i.e., to be able to determine the set of possible worlds in which it is true. This latter set is known in possible worlds semantics as the *proposition* expressed by the sentence.

Possible worlds semantics has been introduced to model in a natural way intuitive explanations of why certain modal sentences are true or false. Intuitively, for instance, a sentence like (a) *It is possible that A* is true, because we can imagine that A would be the case, and (b) *It is necessary that A* is true because we cannot imagine A not being the case. These intuitions are modelled by possible worlds semantics (i) by thinking of a possible world as a possible, or imaginable, way things might have been; how we can imagine that the total state of the world in all relevant aspects could have looked like, and (ii) by counting (a) and (b) as true iff there is a possible world in which A holds, and A is true in all possible worlds, respectively. In other words, whether a modal sentence is true or not depends on the proposition expressed by the embedded sentence.

The framework also allows us to model in a very natural way intuitive explanations for certain puzzles arising with modal discourse. For instance, how can it be that, on the assumption that the number of major planets is 9, we are not simply allowed to *substitute* the description *the number of major planets* for the number 9 in the sentence *It is necessary*

that $(9 > 7)$ without change in truth value? The intuitive explanation is that although the actual number of major planets is 9, we can imagine it being the case that our sun has not 9, but 6 major planets. Thus, it is not part of the meaning of the phrase *the number of planets* that it is equal to 9. Because we can imagine the number of planets being different from 9, in particular because it could be smaller than 7, we evaluate the sentence *It is necessary that the number of major planets* > 7 as being false. In possible worlds semantics this intuition is modelled in the following way: To determine whether *It is necessary that A* is true, we first have to determine the proposition expressed by the embedded clause. According to truth-conditional semantics, we can determine the proposition expressed by a sentence compositionally from the semantic values of its (direct) parts. To be able to do this, we have to know what the semantic values are of their (direct) parts. In particular we have to know what the semantic values are of descriptive noun phrases like *the number of planets*. Given that the semantic value of a sentence, a proposition, is a function from possible worlds to truth values, it is only natural to assume that the semantic value, or meaning, of such a noun phrase is a function taking a world as argument and has an individual, or number, as its value: the unique individual/number that satisfies the description in that world. Because there might be worlds where the number of major planets is less then 7, the statement is predicted to be false.

Until now we have only said when a modal sentence is true in the *actual* world, and looked only at what proposition was expressed by the *embedded* sentence. But implicitly we have also determined the truth conditions under which the *embedding* sentence is true, and thus we have also determined the proposition expressed by this embedding sentence. But if embedding sentences determine propositions, we can also determine when sentences with *iterated* modalities are true. Indeed, one of the most nice features of possible worlds semantics is that it allows for the analysis of sentences with iterated modalities without any complication. But how could iterated modalities ever be interesting? Indeed, they would not be interesting if we would stick with our assumption that a sentence like *It is necessary that A* is true iff A is true in *all* worlds of the model. If all worlds of the model would always be relevant to determine the truth value of the sentence above, the sentence would have the same *truth value* in all worlds. To allow modal statements to be true in one world, but false in another, we introduce an *accessibility relation*, R, between worlds. For our above necessity statement to be true in world w, the embedded sentence has to be true in all worlds that stand in the R-relation to w, $R(w) = \{v \in W |\ wRv\}$, which need not be the same as W, the set of all possible worlds in the model. Because for two different worlds, w and v, $R(w)$ need not be the same as $R(v)$, we can say that what is necessarily true in a world, is a *distinguishing fact* about this world.

From a possible worlds semantics point of view it seems only natural to analyze *belief attributions* in a similar way to how we analyzed modal sentences above. First, what somebody believes is a distinguishing fact about this world, and can be modelled in possible worlds semantics in terms of an accessibility relation. If R_j is John's accessibility relation

and w the actual world, each element of $R_j(w)$ might be the actual world according to John. Note that $R_j(w)$ is a set of possible worlds, a proposition. Second, we can now say that a sentence like *John believes that A* is analyzed as true in w iff A is true in all worlds of $R_j(w)$, i.e., if the proposition $R_j(w)$ is a subset of the proposition expressed by A. What this theory suggests is that the *object of belief* is of the same nature as the *semantic value of a sentence*; both are propositions and modelled by sets of possible worlds.

In this chapter I will defend the claim that, to a large extent, this is indeed the right way to analyze belief, and belief attributions. Following the lead of Stalnaker, I will argue (i) that there exists an independent *philosophical motivation* for modelling the content of belief states in this way, and (ii) that many of the (apparent) problems for this coarse-grained analysis disappear when we take the *context dependence* of belief attributions seriously. But before I will come to these arguments, it is useful to discuss some issues in the theory of reference first.

1.3 The description theory of reference

We have seen above that to be able to determine the semantic value, or proposition, expressed by a sentence compositionally from the meanings of its (direct) parts, we have to know what the meanings, or semantic values, are of these (direct) parts. In particular we have to know what the semantic values are of common nouns and proper names.

According to the traditional theory of meaning, the meaning of a noun or name like 'N' is given by a set of properties or predicates. The conjunction of this set of properties is then called the *meaning* of the expression, and this meaning determines what the expression denotes in the (actual) world, i.e., what its *extension* or *denotation* is. Thus, however the actual world looks like, if P is one of the set of properties that constitutes the meaning of N, the extension of N will always have property P. As a result, being a P is a *necessary* condition for falling under the extension, or denotation, of N.[1] Moreover, if P_1 until P_n is the set of properties or predicates that constitutes the meaning of N, being an individual that has all the properties P_1 until P_n is also a *sufficient* condition for being in the denotation of N.

In terms of possible worlds semantics we might say that the semantic value, or *intension*, of a common noun or proper name is a function from worlds to the individual, or a set of individuals, in that world that satisfies all the properties that together constitute the meaning of the proper name or common noun. Thus, the meaning of the noun *determines the extension* of the noun in each of the possible worlds. If P is a predicate that partly constitutes the meaning of proper name N, the sentence *N is a P* will be an *analytic truth*,

[1] Some proponents of the description theory have weakened the theory; it is not demanded that the individual has *each* of the properties of the set that constitutes the meaning of N, but only that it has *most* of the properties. As a result, being a P is no longer a necessary condition for being an N. But it is, of course, still a necessary condition that an N has *most* of the relevant properties.

in the sense that it will be true in all worlds in which N denotes an individual.

If we restrict our attention for the moment to proper names, we can follow the Fregean tradition and say that the set of predicates that forms the meaning of the name is the *sense* of the name, and that this sense determines the *referent* or denotation of the name, if it has any. This set of predicates forms a description, and thus we might call this traditional theory of meaning for proper names the *description theory of reference*.

This description theory of reference for proper names might be compared with the alternative, *Millian*, theory of meaning, according to which the semantic value, meaning, or intension, of a proper name is simply its actual referent.

From an abstract semantic point of view, there seems to be no good reason to assume that the traditional theory of meaning of proper names is any better than the alternative Millian theory. Both enable us to determine the proposition expressed by a sentence in which a name occurs compositionally from the semantic values of its parts. But there are some well known consequences of the Millian theory that seem problematic, and that suggest that the traditional theory must be on the right track. The (seemingly) problematic consequences of the Millian theory fall into two categories.

The first category of problems is *empirical* in nature; the Millian theory, in combination with a compositional semantics, seems to give rise to some implausible predictions in empirical semantics. First, it gives rise to a *substitution puzzle*. If the semantic value of a sentence, the proposition expressed by it, is functionally dependent on the semantic values of its direct parts, it seems to be predicted that within a Millian theory of meaning two proper names that actually have the same referent can always be substituted for each other without change in semantic value of the sentence in which they occur, because the semantic values of the names are identified with their actual referents. However, it seems that this cannot be the case; although *Hesperus* and *Phosphorus* refer in fact to the same planet, the Babylonians did not think that the names referred to the same object. As a result, although they obviously would agree with us that (1a) says something true, they would not agree with us that (1b) denotes something that is true:

(1) a. Hesperus is Hesperus

b. Hesperus is Phosphorus

The most obvious way to account for the difference in informativity between (1a) and (1b) is to say that the two do not express the same proposition; (1a) will determine a proposition that is necessarily true, while (1b) expresses a proposition that is only contingently true, and thus false in some possible worlds. But this way of solving the problem seems impossible for proponents of the Millian theory of names, because the theory predicts that in both cases the same proposition is expressed. A second problem for the Millian theory is that it is not clear what the proposition could be that is expressed by *negative existential sentences* like *N does not exist*, when 'N' is a proper name. If N indeed has no referent,

it has according to the Millian theory also no semantic value. But how then should we determine the semantic value of the sentence in which N occurs?

The second category of problems is of a more fundamental level. The problem is that it is not clear *why* a proper name has the referent it actually has. The Millian theory of meaning for proper names cannot explain why a is the referent of N, if it is.

Where the Millian theory is said to give rise to the above two kinds of problems, it seems that the description theory of reference for proper names does not give rise to any one of them. First, it seems obvious that the *substitution problem* does not arise. Although the names *Hesperus* and *Phosphorus* actually refer to the same individual, this doesn't mean that they also have the same meaning, intension, or semantic value. Exactly because the two names do not have the same semantic value, a proponent of the description theory argues, they cannot always be substituted for each other without change in meaning. This might only be done with two expressions when they not only have the same reference, but also the same semantic value. Second, negative existential sentences are also not problematic, for there might be proper names that have as their semantic values intensions that have no extension in the actual world.

The description theory of reference can also explain very straightforwardly *why* a proper name have the referent is actually has. According to this theory, we associate with a name a set of properties, or a description, its sense, and the name refers to a particular object *because* in the actual world it is this object that satisfies this description. In other words, according to proponents of the description theory of reference, the two questions (i) *What* is the semantic value of a name?, and (ii) *Why* does the name has the referent it actually has? are interrelated, and should be given a single answer. The Millian theory cannot give a single answer to both of these questions, because it does not even start answering the second one. The description theory of reference, on the other hand, is able to give a unified answer to both questions, and is therefore, or so it seems, to be preferred to the Millian theory.

So far, so good. But now notice that the second question asked in the Millian theory has a somewhat different nature than the second question asked in the description theory. In the Millian theory, reference and semantic value are one and the same, but this is not the case for the description theory. As a result, the question of why a name has the referent it actually has asks for an explanation for why a proper name has the *semantic value* it has in the Millian theory, but not in the description theory.[2]

Thus, it seems that the more fundamental reason why the description theory of reference is to be preferred to the Millian theory is not so convincing as it looked at first blush. It would only be preferred if it allows for a better and more natural answer than the Millian alternative to the question of *why* a proper name has the *semantic value* that it actually has.

Although the description theory of reference cannot give a single answer to this latter

[2]Cf. Stalnaker (1998a).

question and the question relating to *what* the semantic value of a name is, it seems that it is able to give a natural answer to the why-question. It seems reasonable to assume that *referring* is something that *speakers* do with terms, and not something done by the terms themselves. Thus, if a proper name refers to an individual, it is the speaker who refers to this individual by his use of this name. Why does the speaker refer to this individual with his use of this name? Because the meaning of the name is a particular description, or a set of properties, and it is this individual that satisfies this description, or has these properties. But now our new question comes up: Why is *this* particular description the meaning of this name? The natural answer now seems to be that this is the case because the *speaker* associates with the name this description. Thus, in the end, our question *Why does proper name 'N' refer to individual a?* is answered by proponents of the description theory of reference in the following way: because speakers that use the name N (tend to) associate with the name a particular description, and that a is the (unique) individual that satisfies this description. Let us call this resulting account the *description theory of speaker's reference.*

1.4 The description theory of reference and externalism

However natural this description theory of speaker's denotation might be, Donnellan (1970) and Kripke (1972) have shown that it leads to counterintuitive results. Kripke argued that uniquely fitting some set of descriptions that the speaker associates with a proper name is neither a necessary nor a sufficient condition for a successful use of it. It is *not necessary* that *the speaker* has an identifying set of descriptions in mind for the successful use of a proper name, because ordinary people can, for instance, use the name *Feynman* to denote the physicist Feynman even though they have no uniquely identifying set of descriptions in mind. To uniquely satisfy all or most of the descriptions associated with a proper name is also *not* a *sufficient* condition for an individual to be referred to by the name. This point is made clear by Kripke's Gödel example. If someone associated with the name *Gödel* only the description *prover of the incompleteness of arithmetic* he would still denote Gödel and be saying something false of him in uttering *Gödel proved the incompleteness of arithmetic* if somebody different from Gödel was the actual prover of what is known as 'Gödel's incompleteness theorem'.

Besides giving similar kinds of counterexamples to the description theory of speaker's denotation, Donnellan (1970) also pointed out that what is referred to by a proper name by a speaker on a particular occasion depends not only on the intention of the speaker, but also on the conversational context. Consider a student who is known to be acquainted with two different people with the name *J.L. Aston-Martin*. One of these is a famous philosopher whom the student knows from having read some of the books that the philosopher has written; and the other a non-famous person whom he knows from a recent party.

Unfortunately, the student wrongly assumes that both persons are one and the same: that he is acquainted with one person named *J.L. Aston-Martin* in two different ways. He associates with the name a set of descriptions that does not uniquely fit one individual; some descriptions fit the famous philosopher, and others the man he met at the party. Still, Donnellan argues, in some conversational situations the student will unambiguously refer to the famous philosopher by his use of the name *J.L. Aston-Martin*, while in others he will unambiguously refer to the man he met at the party. So, although there is no single individual that satisfies all descriptions the speaker associates with the name, still he might be able to refer with the name to a particular individual. Thus, uniquely satisfying the set of descriptions that the speaker associates with the name, is not a necessary condition for referring with a name to a particular individual. Moreover, Donnellan suggests that which individual he *does* refer to with his use of the name depends crucially on the conversational situation; in particular on (what he assumes to be) the attitudes of the other participants of the conversation.

Although we have suggested in the foregoing section that the description theory of meaning can account for the substitution puzzle, it is important to notice that it is not at all clear that the description theory of *speaker*'s denotation can account for the puzzle (cf. Kripke, 1979). Remember that according to the description theory of meaning two names can be substituted for each other without change in meaning of the clause in which they occur, if we associate with the two names the same description. But now suppose that some agents associate with the names *Cicero* and *Tully* the same description, while others associate with the names two different descriptions. In that case it is not at all clear, according to the description theory of speaker's denotation, what is expressed by the embedded sentence of the attitude attribution *Many are unaware that Tully is Cicero*.

Perhaps the extension of a proper name depends not so much on the descriptions the *speaker*, or the *relevant agent*, associates with it as on the set of descriptions that *most people in the relevant linguistic community* associate with it. It is then this set of descriptions that determines the reference. However, Donnellan and Kripke have shown that this, too, cannot be the case. Kripke's example of Gödel, for instance, shows that this has counterintuitive results. And as Donnellan and Kripke have also observed, if we associate with the name *Aristotle* the description *the teacher of Alexander*, it would also lead to the conclusion that the statement *Aristotle was the teacher of Alexander* is true solely because of the meaning of the proper name. This again seems counterintuitive.[3]

Kaplan (1989) and Perry (1977, 1979) have argued against description theories of reference of *indexicals* (pure and demonstrative) on grounds very similar to those presented by Kripke and Donnellan. There seems to be no plausible candidate for the speaker's meaning of an indexical like *today*. First, it cannot be the description the *speaker* associates

[3]We have seen earlier that some have argued that a set of descriptions fits a unique individual if this individual is the unique individual that fits most descriptions of this set. Donnellan and Kripke also convincingly argued against such a weaker variant of the description theory.

with the relevant day. For suppose that the description I associate with 7 October picks out 8 October, instead of 7 October, because I have taken a long nap. In that case, we would predict that the proposition expressed by my utterance of *Today the weather is fine* on 8 October was that on 7 October the weather is fine. Clearly, this is the wrong prediction. Second, suppose that the meaning of *today* is what a competent speaker of English associates with the word *today*. In that case the intension of *today* does not change from day to day. But this should be the case if the intension of the sentence is compositionally determined by the intensions of its parts, and the proposition expressed by it depends on the day of utterance. Similar arguments can be given against a description account of pronouns demonstratively used.

By very much the same kind of arguments, Kripke, Putnam and Burge have convincingly argued that the set of properties that speakers or agents associate with *common nouns* should also not be equated with the meaning of the noun. First, the meaning cannot be the description the *speaker* associates with the term. This is made very clear by the 'Twin Earth' stories given by Putnam (1975) and Burge (1979). These stories always involve a comparison between two almost identical persons (twins): one in the actual world and one in a counterfactual world, Twin Earth, minimally different from the actual world. In Putnam's story, the stuff that the inhabitants of the counterfactual situation call *water* is superficially the same as the stuff *we* call *water*, but its chemical structure is not H_20, but XYZ. If, then, both the earthling and his twin assert *Water is the best drink for quenching thirst*, intuitively they have said something different. But how can this be if they associate exactly the same description with the word and if speaker's description determines reference? A similar 'Twin Earth' story invented by Burge (1979) shows that the problem is not limited to a small set of terms. In fact, stories can be invented for almost any expression to show that it is not the description that the speaker associates with an expression that determines its extension. The reason is that the linguistic practices of members of the agent's community are crucial in determining the extension of a term. Perhaps what counts, then, are the properties associated with the term by most speakers, or the relevant specialists of a linguistic community. But Putnam shows that this cannot be the case for natural kind terms either. The demonstration involves the same 'Twin Earth' story, but now set in 1750. Specialists on Earth and Twin Earth are not yet able to see any difference between H_20 and XYZ. But intuitively, even if a typical Twin-Earthian (twin-) English speaker utters *Water is the best drink for quenching thirst* on Earth, he is not talking about H_20.

On the basis of these arguments Kripke and Putnam claim that the meaning of at least proper names and natural kind terms is not the set of descriptions associated with them, but simply what they refer to. But then the question arises of *why* the proper name or natural kind term refers to this particular entity or stuff. At this point, Kripke (1972) and Putnam came up with their *externalistic* answer. This externalistic answer has both a social, and a causal aspect. The *social* aspect demands that to use N to refer to a,

there must be a convention among the speakers of the relevant linguistic community that N could be used to refer to a. For proper names Kripke (also?) demands the following *causal* condition: Proper name N can refer to a, only if, and because, a is the entity that is the *source* of the reference-preserving link from the initial baptism of the expression to the speaker's use of the name.

It is sometimes assumed that the lesson to be learned from the criticism of the description theory of speaker's denotation is that what one refers to with an expression is not dependent on what the speaker believes and intends, but only on causal and social conditions *external* to the agent. However, this is not exactly the moral Kripke (1972) drew from his discussion. He argued explicitly that the existence of a causal link by itself cannot be enough, since it leaves out an important *intentional* element. It should at least also be the case that the speaker intends to use the expression in the same way as it was transmitted to him via other members of the community. If this is the way the externalist theory should be understood, a speaker cannot only refer with a name to a particular entity, but he can also *intend* to refer to this particular entity. How can someone intend to refer to a particular entity by his use of an expression, if he does not associate a particular description with this expression which can only apply to this particular entity? Can someone not only intend to refer to a, if he has *beliefs about* a, and how can one have beliefs about a if one has no information to identify a?

First note that intuitively you can have beliefs about a, even if you cannot identify a in all possible circumstances, i.e. if you have no *eternal* description in mind that satisfies a. This suggests that aboutness should not be cashed out in terms of 'having a description in mind by means of which you can indentify the relevant object in all possible circumstances'. Proponents of the causal information-theoretic account of intentionality propose, instead, that not only the aboutness relation between expressions and objects, but also the aboutness relation between belief states, and other information states, and objects should be explained in terms of causal relations.[4] In this way they can account on the one hand for the natural assumption that what one says and refers to depends on what one intends and believes. But because what one has intentions and beliefs about should be explained, according to this account, in terms of causal and other externalistic relations, we need not have to make the problematic assumption that proponents of the description theory of speaker's denotation make; that noun N refers to a because the speaker associates with N a set of descriptions, and that a is the unique individual, or set of individuals, that satisfies these descriptions.

Evans (1973) was perhaps the first to propose that the causal theory of reference should be based on a causal theory of belief, or of information. He argued with Kripke that a causal link for proper names is necessary, but that this causal link should not be between the initial naming and the speaker's current use of the name, but rather between the *body of information* relevant to the speaker's use of the proper name on a particular occasion

[4]See section 1.6 for more about this.

and the object that is the dominant causal origin or source of this body of information.[5]

That it should be the *dominant* source is important to account for the case where two, or more, objects are the source of a particular body of information relevant to the speaker's use of a referential expression on a particular occasion. In those circumstances, according to Evans, we say that on that occasion the speaker refers with the expression to the object that is the *dominant* source of this body of information. By making the referent of a name dependent on the dominant source of the relevant body of information, Evans can account for the fact that certain names, like *Madagascar*, have changed their denotation. Consider the following example:

> If it turns out that an impersonator had taken over Napoleon's role from 1814 onwards (post Elba) the cluster of the typical historian would still be dominantly of the man responsible for the earlier exploits and we would say that they had false beliefs about who fought at Waterloo. If however the switch had occurred earlier, it being an unknown Army officer being impersonated, then their information would be dominantly of the later man. They did not have false beliefs about who was the general at Waterloo, but rather false beliefs about that general's early career. (Evans, 1973, p. 202)

With a referentially used expression we refer to the dominant source of the information 'responsible' for that use of the expression on a certain occasion. In sum, according to Evans' informational account of proper names, the information associated with a proper name plays its part, although the causal link is necessary. Since this causal element is still part of the analysis, a is the referent of proper name N not because a fits best with the information associated with N, but because it is the dominant source of this body of information. An object can be the dominant source of a particular body of information even if it does not fit this information very well. It follows that if P is one of the properties we associate with N, we still do not know that the sentence N *is* P is true by necessity. This causal information-theoretic theory of aboutness can also account for the above observed phenomena regarding the Twin Earth examples. I will come back to this in section 1.6.

For a speaker to refer with a proper name to a particular entity on a particular occasion, it is not enough that this entity be the dominant source of the information of the *speaker* that is relevant for his use of the name.[6] For N to be a name for a it should also be the case that there exists a convention among the speakers of the relevant linguistic

[5]For more on the distinction between purely causal account and informational accounts, see Dretske (1981, ch. 1). Kripke (1972), (addendum (e)) argued that Evans was really a proponent of the description theory after all, by his use of the notion of 'information'. At least since Dretske (1981) it has been clear that Evans was not.

[6]But it is a necessary condition for a speaker to refer with N to individual or stuff a that a be the dominant source of the content of *his thoughts* relevant to his use of N. How else to account for the intuition that when a typical Twin-Earthian (twin-) English speaker utters *Water is the best drink for quenching thirst* on Earth, he is not talking about H_2O?

community that N could be used to refer to a. This convention brings in the *social element* of the meaning of a proper name, and is to be explained in terms of beliefs and intentions of the members of the community. Thus, the analysis of proper names here sketched is in *two* ways externalistic: (i) it is externalistic because the *causal* relation is an external relation, and (ii) it is externalistic due to the *social* part of the story.

The social aspect of the theory is needed, for instance, to account for the fact that the student Donnellan (1970) talks about can refer to two different individuals by his use of *J.L. Aston-Martin* in different conversational contexts. In the two different conversational contexts that we are in, there are different bodies of information that the student has which are relevant, and these have different sources. In some conversational contexts it is a convention that the name *J.L. Aston-Martin* can be used to refer to the famous philosopher, in others it is a convention that it can (also) be used to refer to the man the student met at the party last night. In these different circumstances it might be different bodies of information that are relevant for determining the referent.

1.5 The pragmatic account of intentionality

The problem of intentionality is to explain how certain things can represent, be about, or be directed to, other things. A number of things have this representational capacity; natural language and minds or attitude states of agents are two prominent examples among these. According to one account of intentionality, the representational capacity of language and other media should be explained in terms of the intentionality of the attitude states of agents. So, how can an attitude state represent something 'outside of itself'? The usual (holistic) strategy for answering this question is to say that this representation relation has to be explained in terms of relations to propositions. An agent represents or is directed towards another physical object if he/she stands in a certain attitude relation to the proposition that involves or is defined in terms of this other physical object. But then, what is it to stand in a certain attitude relation to a proposition? Propositions are abstract objects, and how can we stand in a relation at all to such an abstract object? Measurement theory suggests an answer: an agent stands in a relation to an abstract object if by this object we can measure the state of the agent. For a dispositional predicate like *believe*, this measure is usually stated in terms of counterfactual relations.[7] That is, agent x stands in a relation R to the proposition $P, R(x, P)$, if there is a certain predicate F such that the following counterfactual is true: if it were the case that P, then x would be F (Stalnaker, 1984, 1994). Because F is a predicate saying something about how x stands in the world, a true attitude attribution characterizes an individual as being in a certain state by saying something about the relation the individual bears to the world.

According to a purely *pragmatic* (and almost behaviouristic) account of intentionality,

[7] By 'counterfactual' I mean here, and in the rest of this book, the kinds of conditionals that are analyzed in Lewis (1973).

as assumed by most proponents of decision- and game theory, and explicitly defended by, for instance, Ramsey (1931), Dennett (1969), Stalnaker (1972), and Lewis (1974), the R is explained in terms of something like a *tendency to bring about* relation. An attitude of an agent is about, or directed to, an object or state of affairs because the agent is disposed to perform actions that involve this object or state of affairs. The attitude of the agent is this dispositional state. Proponents of this account assume that attitude attributions are normally made to explain behaviour. A person performed a certain action because by doing so he could satisfy his desires in a world in which his beliefs are true. For this to work, it has to be presupposed that the behaviour of agents can be rationally understood; and that attitudes, in particular belief and desire, are correlative dispositions, or *functional* states, of such a rational agent. These states are individuated by the role they play in determining the behaviour of the agent who is in such a state. For instance, if R is *desire*, for an agent to desire that P means that the agent is disposed to perform actions that tend to bring about P in a world in which his beliefs were true. Analogously, if R is *belief*, the agent believes that P, if he is disposed to perform actions which tend to satisfy his desires in worlds in which P and his other beliefs were true (Stalnaker, 1984, p. 15). A belief/desire system is correctly attributed to an agent if the actions that the agent performs can be explained in terms of this belief/desire system (and a theory of rational behaviour). And if the actions of the agent that involve object a can be explained in terms of the intentional state, the intentional state of an agent can be said to be about object a. So, in a manner similar to the description theory of reference, aboutness is determined by *fit* with reality.

The pragmatic picture suggests that propositions, the objects of attitudes, should be modelled by sets of possibilities. Why? The reason is that rational agents are seen as *deliberators*. A deliberator is an agent who considers various possible actions and determines his choice by his beliefs about the possible outcomes of these alternative actions and by the desirability of these possible outcomes. This picture of rational activity suggests that the primary objects of attitudes are sets of alternative outcomes of possible actions, or alternative ways that the world might be (Stalnaker, 1984, p. 4). These possible outcomes of actions can be thought of as possibilities that are maximally specific with respect to all of the issues relevant in the deliberation. Thus, if we want to say that an attitude state that represents the beliefs of an agent is modelled by a set of possibilities, these possibilities are only as fine-grained as demanded by the conversational context.

Not only should belief states be modelled by sets of possibilities, but so should all kinds of acceptance attitudes of an agent — in particular, for the propositional attitude of *presupposition* (Stalnaker, 1973, 1974), approximately the attitude of common belief (cf. Lewis, 1969). Just as beliefs and desires are functional states of rational agents, so too are presuppositions. All these attitude states are relevant to a theory of rationality, which is required to explain why certain behaviours of rational agents are appropriate when they are (Stalnaker, 1972). The attitude of presupposition, or common belief, is needed to explain the agent's behaviour when he is engaged in a game with other participants. To

explain certain actions of the agent in a game, it is normally not enough to assume that the agent knows the rules of the game, or that he assumes that everybody will maximize his own expected utility. The agent has to assume something much stronger: that these things are presupposed, or commonly believed. One of the games in which a rational agent can participate is the communication game. The appropriateness of the communicative actions of the agent in the game is to be explained in terms of what he presupposes.[8]

Saying that the objects of attitudes are sets of possibilities does not necessarily mean that an agent's belief state should be represented by one set of possibilities. It might be that various thoughts the agent has are not integrated. In these cases he doesn't have a single coherent conception of the world. As a result, a belief state should not be modelled by a set of possibilities, but rather by a set of such sets. Each compartment can be thought of as what he would believe if a certain question were asked or a certain problem posed. Thus, what somebody believes can be thought of as a function from problems to sets of possibilities. This latter set of possibilities is then what the agent *explicitly* believes, how he is disposed to act, in a situation in which the problem is posed. What somebody *implicitly* believes might be thought of as the union of the set of compartments. Although implicit beliefs are closed under logical consequence, for explicit beliefs this is the case only with respect to each compartment separately. Stalnaker (1984, ch. 5) suggests that deductive reasoning might be thought of as trying to integrate different compartments of one's belief state with each other.[9] The beliefs of agents are not closed under deduction because not every compartment is always accessible. Questioning helps to make explicit what was only implicitly believed before (Stalnaker, 1991).

1.6 Intentionality: the causal/informational account

Although the account of intentionality sketched above is essentially an externalist one, we have seen already in section 1.4 that the arguments of Kripke, Putnam, and Burge that are problematic for the description theory of meaning are also a threat to the purely pragmatic account of intentionality.[10] The pragmatic account of intentionality alone leaves the content of belief underdetermined. It cannot explain why the Earthling Oscar (one of the twins in Putnam's Twin Earth example) is *thinking about* H_2O, if he is thirsty and asks *Can someone give me some water?* According to a purely pragmatic account of intentionality, he might as well be thinking about XYZ. But we don't need these artificial Twin Earth stories to see that the pragmatic account by itself cannot solve the problem of intentionality. Just like for the description theory, *fit* is not enough.

[8] See chapters 2 - 4 for more on this.

[9] See also the neighborhood semantics of Montague (1974), and the cluster models of Fagin & Halpern (1988).

[10] Although they were, in the first place, directed against individualistic accounts of intentionality.

> What makes an assignment of a system of belief and desire to a subject correct cannot just be that his behaviour and behavioural dispositions fit it by serving the assigned desire according to the assigned beliefs. The problem is that fit is too easy. The same behaviour that fits a decent, reasonable system of belief and desire also will serve countless very peculiar systems. Start with a reasonable system, the one that is in fact correct; twist the system of belief so that the subject's alleged class of doxastic alternatives is some gruesome gerrymander; twist the system of desire in a countervailing way; and the subject's behaviour will fit the perverse and incorrect assignment exactly as well as it fits the reasonable and correct one. Thus constitutive principles of fit which impute a measure of instrumental rationality leave the content of belief radically underdetermined. (Lewis, 1986, p. 38)[11]

Just as the arguments of Kripke, Donnellan, and Putnam motivated a causal theory of reference, the Twin Earth stories of Putnam and Burge motivated a causal theory of intentionality. In fact, as we have suggested already in section 1.4, it seems that the causal theory of reference *presupposes* a causal theory of intentionality; how else to explain that speakers can not only refer, but also *intend* to refer to a particular individual without having an identifying description in mind (cf. Stalnaker, 1998a). Some have concluded from Putnam's Twin Earth story that both the pragmatic account of intentionality and the description theory of meaning are generally satisfactory, their problems being limited to only a narrow range of cases. However, as Burge's (1979) extension of Putnam's Twin Earth story made clear, the problem for the above theories is much more general. This suggests that the notion of intentionality should be analyzed, at least partially, in causal information-theoretic terms. How should this causal information-theoretic account be cashed out?

I believe that this should be done in terms of counterfactual relations. Remember that the relation between agent x and abstract proposition P was in general defined in terms of a counterfactual definition of $R(x, P)$. The pragmatic picture explains this relation in terms of a *tendency to bring about*. According to the information-theoretic account defended by Stampe (1977) and Stalnaker (1984), this relation is analyzed in terms of a *tendency to carry information*, a relation of *indication*. For a certain mechanism to be a representational mechanism about a certain environment, the mechanism must be able to be in various alternative states that tend to vary systematically with variations in the environment. When this mechanism tends to be in state R when P is the case, this mechanism's being in state R can be said to contain the information that P; that is, it indicates that P is the case. Normally, the internal state of a representational mechanism tends to vary systematically with the states of the environment, and thus indicates something about the environment; this is because the mechanism's being in a certain internal state is *caused* by this environment. If this is the right way to explain why mechanisms can represent

[11] See also Stalnaker (1984, ch. 1).

something outside themselves, and thus have content, it suggests that the notion of content, or the indication relation, should be analyzed in terms of nested counterfactual conditionals. That is, *if* conditions are normal or optimal, and *if* various alternatives to P were true, then the believer would be in various alternative states (see Stampe (1977) and Stalnaker (1988)). To the normal, fidelity (Stampe, 1977), or channel conditions (Dretske, 1981) belong both conditions external to the agent and those related to the internal functioning of the representational mechanism. The reason for this use of nested counterfactuals should be obvious: if conditions are not normal, P might hold without the internal mechanism being in state *R;* and it is the relevant alternative states that the environment could be in that determine the content of the internal state.

If content is explained in this way, we can explain why Oscar has beliefs *about* H_2O and why Oscar talks *about* H_2O if he uses the term *water* in Putnam's (1975) Twin Earth story.[12] Some of his beliefs are about H_2O because normally these beliefs are sensitive to facts about H_2O; normally he would not express his belief by saying that there is water in the bathtub, if it was not H_2O that was in the bathtub. And he talks about H_2O if he uses the word *water* because the content of one's utterances should be explained in terms of the content of one's representations. But how come that these internal representations are sensitive to facts about H_2O? This is partly so, because we associate with the term *water* certain superficial properties, and normally it is only H_2O that has those properties. Oscar's twin is not thinking or talking about H_2O because on Twin Earth ideal conditions are different; it is not H_2O but XYZ that normally has the relevant superficial properties, and thus it is not H_2O but XYZ that is 'responsible' for the use of the word *water* by Oscar's twin.[13] To account for the intuition that Oscar and not his twin is thinking about H_2O, we have to assume that normality/optimality conditions are determined as is normal/optimal for *us* in the actual world. We will see later, however, that sometimes the relevant normality/optimality conditions can be set in a different way: how they are determined depends on the conversational context. According to the above account, the indication relation is context-dependent not only because of the variability of the normality/optimality conditions, but also because of its dependence on the conversational context. We will see later how this double context dependence can be used to account for some puzzling consequences of the information-theoretic approach to content.

Just as the pragmatic analysis of intentionality motivated a coarse-grained account of propositions in terms of sets of possibilities, any externalist strategy working with a notion like *indication* will also motivate a rather coarse-grained conception of content, the object

[12]See especially Stalnaker (1993).

[13]But then, why is it that when water is H_2O, it is also *necessarily* H_2O? The reason is that the notion of normal or optimal conditions is a modal notion. Once we have found out the optimality conditions in the actual world with respect to the word *water*, these conditions determine a set of worlds in which they hold. In our case we can say that water is necessarily H_2O, because for necessity we look only at worlds in which the optimality conditions in the actual world with respect to the word *water* hold (compare van Fraassen, 1977, and Stalnaker 1993).

of (holistic) attitudes like belief. In particular, it will not allow for a distinction between propositions that have equivalent truth conditions. The reason is that such propositions will behave identically in causal and counterfactual constructions (Stalnaker, 1994). This suggests that the possible world analysis of content is the correct one.

According to Stalnaker (1984, ch.1), the causal account of intentionality should not replace the pragmatic account, but should only complement it.[14] He argued that the causal backward-looking account should take care of the *content* of beliefs, while the pragmatic forward-looking account should take care of the *functions* that different attitudes have in determining or explaining action.

1.7 Combining the pragmatic and causal accounts

Stalnaker argued for a combination of the pragmatic and the causal information-theoretic accounts of intentionality. At first blush this seems to be impossible. According to the pragmatic picture, we can have false beliefs; and we can have, on a particular occasion, the belief that *P,* although on this occasion this is not caused by the information that *P*, but by the information that *Q*. Doesn't this show that the two accounts are incompatible? I don't think so. The information-theoretic analysis of content sketched above, which appeals to counterfactual dependencies and normal conditions, makes no distinction between information and misinformation; rather, the analysis of belief is based on what is *normally* correct. False beliefs are just deviations from the norm. Also, it is crucial that, at least in general, we do not account for content in terms of what was *actually* the cause of a certain information state, but in terms of what *normally* causes this mechanism to be in that state. If normally *Q* does not cause the representational mechanism to be in a state corresponding with *P,* the fact that the information that *Q* actually caused the representational state to be in the particular state that it is in does not demand that the state has content *Q*.[15]

Stalnaker (1984) defends his combined account as a reductive *naturalistic* analysis of intentionality. But in order for this account to be fully successful, it is crucial to explain the normal or ideal conditions in terms of which content is defined in non-intentional ways. I am not sure whether this can be done. What the analysis of intentionality certainly will do, though, is to explain notions that are thought to be mysterious, like *content* and *intentionality*, in terms of notions that are considered not to be so problematic.

Still, it seems that the causal account of content leads to unsolvable problems even if the above problems can be accounted for. Once we accept that the content of expressions and intentional states causally depends on external conditions, we are confronted again

[14] Stampe (1977), Dretske (1981) and Evans (1982) come to basically the same conclusion.

[15] Although I believe that actual cause in general need not be relevant for the analysis of aboutness, I do believe it is for the aboutness relation between a belief state and a particular individual. This causal relation, however, need not be such a strong notion of acquaintance that Kaplan (1969) appealed to.

with many old problems. How can agents seriously believe (doubt) what is expressed by statements whose propositions are necessarily false (true)? On the account just defended, we can no longer account for the fact that we appropriately attribute to agents beliefs that seem to be necessarily true or false. Perhaps the most serious problem that it gives rise to is this one: if we accept externalism, then it seems that attitude ascriptions can no longer do the job commonsense psychology tells us they do. A common sense explanation of why the Earthling and his counterpart drink so much of the stuff that in their respective communities is called *water* if they are thirsty is that they think that what they call *water* is the best drink for quenching thirst. The problem is that according to the causal conception of content it seems that the belief attribution *Oscar believes that water is the best drink for quenching thirst* is more specific than we want, because we know that Oscar cannot distinguish H_2O from XYZ. Any causal, or information-theoretic account of content seems to predict a *too specific* notion of content in these cases.

On other occasions the predicted contents seem *not* to be *specific enough*. The fact that the externalist position leads to an insufficiently specific notion of content can perhaps best be illustrated with sentences in which so-called *essential indexicals* occur. The following example is from Kaplan (1989) and Perry (1979). Kaplan is looking at a mirror and sees a man whose pants are on fire. This man is actually Kaplan himself, but he does not realize this, and stays cool under the situation. After a while things get hotter, however, and he starts to realize that he has been looking at *himself* in the mirror. His earlier coolness disappears, and he shouts *Help, my pants are on fire!* How can this change in behaviour be explained if the first person possessive would refer simply to Kaplan?

If content is determined as is predicted via the causal account, it would seem that completely rational agents could have *inconsistent beliefs*. Consider Kripke's (1979) case of Pierre. Pierre grows up unilingually in Paris and learns something from his parents that is expressed by saying *Londres est jolie*. On this basis, he is inclined to assent to this sentence. On the basis of the disquotation principle[16] and the assumption that meanings are preserved under translation,[17] it seems that we can conclude that the sentence *Pierre believes that London is pretty* is true. Later Pierre goes to England, learns English, settles in an ugly part of London, but he does not realize that the city that he learned about in Paris is the city that he lives in now. He is disposed to utter or assent to *London is not pretty*. By the use of the disquotation principle it seems that we may conclude that the sentence *Pierre believes that London is not pretty* is true. On the assumption that Pierre does not give up his earlier belief expressed in French by *Londres est jolie*, it is hard to see how we can escape the conclusion that Pierre has inconsistent beliefs if the extension of a proper name exhausts its meaning. This is paradoxical, for Pierre may be a perfect logician.

[16] If a normal English speaker, on reflection, sincerely assents to 'p', then he believes that p.

[17] If a sentence in one language expresses a truth in that language, then any translation of it into any other language also expresses a truth (in that language). (Kripke, 1979).

The causal theory of reference seems not only problematic for some belief attributions, it also seems to give rise to insuperable problems for normal conversations. If the communicative actions of rational agents are to be partly explained in terms of what they presuppose, it seems natural that any agent wants to assert only something informative. An assertion is informative only when the acceptance of the proposition expressed by the relevant utterance eliminates some (but not necessarily all) possibilities representing the speaker's presupposition. Now note that if proper names are interpreted in a Millian way, we have to *presuppose* that the names we use have a referent. Because the referent, or extension, of such terms exhausts their intension, no proposition can be determined if the expression has no referent. But this gives rise to a new problem: how can we appropriately use statements by which we assert that a proper name has no (or an empty) extension? Another problem is that it is no longer clear why we sometimes make claims that in our world are necessarily true or false. Normally, a claim makes sense only if it states a contingent proposition. By assuming that for certain expressions the actual extension the expression has exhausts its intension, it also seems to be impossible to explain why certain sentences are always true simply because of the way the words in them are used.

To take an example from Kripke (1972), given that stick S is used to fix the referent of the term *one meter,* we know by definition that the sentence *Stick S is one meter long* is true. Still, the sentence is not necessarily true; we can imagine that the stick is longer than it actually is. How can we account for this intuition? Other examples mentioned by Kripke (1971) and Evans (1979) are the use of names like *Jack the Ripper* and *Deep Throat.* If *Deep Throat* is used as the name for the person in the White House, whoever it was, who was the source of Woodward and Bernstein's Watergate information, how can we account for the fact that the sentence *Deep Throat is used as the name for the person in the White House who was the source of Woodward and Bernstein's Watergate information* cannot be false? The problem is that although the statement above is, in some sense, necessarily true, the intension of the name depends only on its actual extension, which we don't know.

As it happens, the most obvious difficulty we have arrived at will give us an obvious way to resolve at least some of the above puzzles. Although the assumption that the actual extension of certain expressions exhausts their intension can help us to account for the fact that a sentence like *I didn't have to be here, you know* can be true, we can no longer account for the sense in which the statement *I am here, now* is always true.

1.8 Context dependence: two-dimensional semantics

What we have missed until now, of course, is the insight that the extension of an expression depends not only on the situations in which its meaning is evaluated (the index), but also on the *context* in which the expression is used. What is expressed by a sentence is *context-dependent*, so in different contexts the same sentence can express different propositions.[18]

[18]For an extensive discussion of theories of context dependence, see Zimmermann (1991).

Strawson noted this already in his criticism of Russell's description theory, but he concluded that sentences that are context-dependent cannot be handled by formal means. This was overly pessimistic: context dependence, it turns out, can be handled formally if one recognizes the importance and distinct roles of context and index. The context partially determines *what* is said, but does not evaluate whether what is said is *true*; while the index evaluates only the *truth value* of what is said. So modal logic should be sensitive to *pairs of situations*, instead of only single situations. The need for and possibility of a formal treatment of context dependence by means of a separation of the roles of context and index was recognised by a number of people at about the same time.[19] This liberalization of modal logic has proven to be increasingly important in the philosophy of language and in natural language semantics. In Kaplan's (1989) theory of context dependence, contexts and indices are entities of different kinds. A context, c, consists of certain aspects of a world, like speaker, hearer, time, etc. For some cases of context dependence, a world also has to be an element of a context. What makes propositions true or false are worlds, and these are accordingly called indices.

Besides helping to solve some of the puzzles discussed above, there are two reasons why the distinction between context and index is important. The first reason is that in this way we can explain why there are two ways people can disagree about the truth value of a statement (see Stalnaker (1978)). Suppose that the speaker claims something by uttering a sentence, and the hearer disagrees. They can disagree because the hearer has *misunderstood* the speaker. The hearer has made a wrong guess about the context of utterance the speaker was in, and thus about the context-dependent proposition expressed by the speaker. It is also possible that they agree about what is said, but *disagree about the facts that determine the truth value* of what is said. The second reason the distinction between context and index is important is that the distinction makes it possible to handle context-dependent expressions in embedded contexts in a compositional manner without relying on the predicate-logical notion of scope (see Kamp (1971)). Because in normal situations the context of utterance and the point of evaluation of a sentence are the same, it would seem that words like *now* and *actually*, generally speaking, are superfluous. But they are not, as their occurrence in embedded clauses shows. In the following sentences we cannot leave the indexicals out without a change of meaning:

(2) I learned last week that there would *now* be an earthquake.

(3) I would like to have more money than I *actually* have.

If two situations are relevant for determining the truth value of a sentence, we might say that the meaning of a sentence is a relation between two situations, a two-dimensional intension. Following Kaplan, we can call this kind of meaning the *character* of a sentence. The character of a sentence is compositionally determined by the characters of its parts.

[19] For a short history of the subject, see van Fraassen (1977).

If E is an expression, we might call $[E]$ the character of E. Given a context, c, $[E](c)$ is the *content* or *intension* of E. $[E](c)(i)$, finally, is the *extension* of E, if i is an index. The content of a sentence is a proposition, and its extension a truth value. To determine the intensions of (2) and (3) in terms of the intensions of their parts, we have to determine the intension of their embedded sentences with respect to the context of utterance. Double indexing is needed for reasons of compositionality, if words like *now* and *actually* are treated as singular terms or as one-place propositional connectives.

In Kaplan's two-dimensional theory of meaning, a context is something like a tuple that contains an agent, a time, a place, and perhaps a world. A character is a kind of meaning, and is associated with a *type* of expression in a certain language. Only the intension of indexicals and demonstratives is assumed to be context-dependent. But, as proposed by van Fraassen (1977) and Haas Spohn (1994), we might generalize Kaplan's notion of a character such that also the intensions of other expressions become context dependent. Where Kaplan assumes that the character of *I* is not its referent, but rather a *function* from a particular context that contains a speaker to the speaker of that context, van Fraassen proposes that the character of, for instance, a proper name like *Hesperus* is not the object (or better, constant function from indices to this object) that speakers of English refer to by their use of the term, but rather a *function* from a particular context that contains a *distinguished language* that is spoken, to the object that speakers of that language refer to by their use of the term. Similarly, as proposed by Haas Spohn, the character of *water* can be thought of as a function from a context that contains a distinguished language and a world, to the stuff that speakers of that language refer to by the word of that context.

In classical one-dimensional modal logic, there can be only one kind of rigidity and one kind of necessity. But when we make a distinction between the role of contexts and the role of indices, we can make a distinction between *three kinds of rigidity*, *three kinds of necessity*, and *three kinds of entailment relations*. However, abstracting away from all possible languages, there won't be a lot of interestingly new examples of rigid expressions, necessary statements, or entailment relations between expressions. However, we might restrict our contexts that we quantify over. For example, because every context contains a distinguished language, we can determine the class of contexts where a particular language is spoken; the set of contexts where the distinguished language is this particular one.[20] In this way we might single out the class of contexts where English is the distinguished language. Once we have singled out the class of contexts where English is spoken, we can relativise the notions of rigidity, necessity and entailment with respect to this class.

[20]But we have to be careful here. If $\langle a, t, p, l, w\rangle$ is now a context, at first sight it seems that l is the distinguished language of the context. Now suppose that l is English, w the actual world, and w' Twin Earth in Putnam's story. Then the intension of *water* is H_2O in $\langle a, t, p, l, w\rangle$, and XYZ in $\langle a', t', p', l, w'\rangle$, just as desired. But, as we have seen in section 1.4, even if Oscar's twin utters *Water is the best drink for quenching thirst* on Earth, he is intuitively not talking about H_2O, but about XYZ. To account for this, Haas Spohn (1994) proposes that Oscar and his twin do not really speak the same language, in particular, that Oscar's twin does not speak English.

Let's begin with rigidity. An expression has a *rigid content* if it has a constant content in each context. Indexicals like *now*, *actually* and *I* are of this kind in English. If it has the same content in all contexts, it has a *rigid character*. Finally, an expression is *superrigid* if it has both a rigid character and a rigid content. In English, the logical connectives can be taken to be examples of this kind of expression. From now on we will refer to an expression with a rigid content as simply *rigid*, and one with a rigid character as having a *constant character*. Now the three kinds of necessity. First, what a sentence expresses in context c can be true in every relevant world, $[A](c) = K$, where K is the set of all relevant worlds. Sentences like *Hesperus is Phosphorus* and *I am Robert* are necessary in this way, because proper names and indexicals have rigid contents. This kind of necessity is sometimes called *metaphysical necessity*. Second, a sentence can be true in every context in which it is expressed. If $i(c)$ gives us the world of c, this means that for all $c : i(c) \in [A](c)$ holds. Some have identified the necessity of such sentences with that of *a priori necessity*. A sentence like *I am here now* is a well-known example of this kind. If a sentence expresses in every context a proposition that is true in every world, the sentence might be called *analytically true*, the third kind of necessity. According to Haas Spohn (1994), English sentences like *Tulius is Cicero* and *Every ophtalmologist is an oculist* are of this sort.

Now we can define three kinds of entailment relations. First, the *classical entailment* relation: B follows from A in c, $A \models_c B$, iff $[A](c) \subseteq [B](c)$.[21] Second, something that might be called *diagonal entailment*: $A \models_d B$ iff $\forall c : \; i(c) \in [A](c) \rightarrow i(c) \in [B](c)$. The strongest notion of entailment might be called *analytic entailment*: $A \models_a B$, iff $\forall c, w : \; w \in [A](c) \rightarrow w \in [B](c)$. Note that if $A \models_a B$, then both $A \models_c B$, for all c and $A \models_d B$ follow. It should be obvious that just as the notions of necessity can be relativised to particular classes of contexts, so can the notions of entailment.

In this book, or at least in this chapter, I will normally not stretch the Kaplanian notion of a character in the same way as suggested by van Fraassen and Haas Spohn. Instead, I will follow Stalnaker (1978, 1981, 2001), who works in a two-dimensional modal logic, making use of a *token analysis*.[22] Where Kaplan proposed that we should associate two-dimensional objects, characters, with *types* of expressions, Stalnaker proposes that we can associate two-dimensional objects with *tokens* of expressions. He calls these two-dimensional objects associated with sentential tokens *propositional concepts*. Where a character of a sentence is a function that takes a context as argument and has the proposition expressed by this sentential type in this context as value, the propositional concept that can be associated with a sentential *token* is a function from possible worlds to the proposition expressed by this sentential token in this world.[23] We see that making use of a

[21] I am leaving out quantification over models here.

[22] For a discussion of the difference between the *semantic* interpretation of the two-dimensional framework as proposed by Haas Spohn and others and the *metasemantic* interpretation as used by Stalnaker, see Stalnaker (2001).

[23] It seems that the Stalnakerian analysis makes Kaplanian characters superfluous. That's not the way

token analysis, allows us to assume that although context and index have a different function, they still can be entities of the same kind, possible worlds. That's why this analysis is sometimes called *two-dimensional modal logic*.[24] It's instructive to look at a particular example.

Suppose that being engaged in a conversation with Hans and Ede, Ulrike says *I learned a lot from you about the analysis of indexicals*. Obviously, the proposition that Ulrike expresses depends on who the addressee is, which in turn depends on the intentions related to her use of *you*. So, if we know enough about Ulrike's intentions, we can determine what proposition she has expressed. But now the question arises of how we can determine which *propositional concept* should be associated with her token of the sentence. It is easy to determine the propositional concept associated with the sentential token for a certain limited set of worlds. What Ulrike says should be interpreted with respect to a context that represents what is common ground between the three engaged in the conversation, and thus presupposed by them. This context is represented by the set of worlds that, as far as they presuppose, might be the actual world. On the assumption that Hans and Ede have heard what Ulrike has said, she can assume that at the moment that she has made her utterance, she can also presuppose that she has made her utterance. This means that her sentential token does not only exist in the actual world, but also in each of the worlds consistent with what they presuppose. Now we can determine in each of those worlds which proposition she would have expressed, if that world were the actual world. For each world in this set the referent of the token of *you*, and thus the proposition expressed by the sentential token, will depend on Ulrike's intentions in that world, which need not be the same as what she actually intends.

Let us assume with Stalnaker that a world determines both the proposition expressed by a token of a sentence and the truth value of what is expressed. In that case we might say that, if $\underline{A}$ is a sentence token of A, $[\underline{A}]$ is a relation between worlds. Let's say that $w[\underline{A}]w'$ means that what is expressed by $\underline{A}$ in w, $[\underline{A}](w)$, is true in w'. Let $K(w)$ be the set of all relevant possible worlds, the set of worlds that represents what the participants of the conversation presuppose in w, the actual world. The proposition expressed by $\underline{A}$ in w is, of course,

$$[\underline{A}](w) \quad = \quad \{w' \in K(w) |\ w[\underline{A}]w'\}$$

This proposition is known as the *horizontal proposition* expressed by $\underline{A}$ in w. For another important proposition, we introduce the diagonal operator '†', the *dagger*, in the following way:

---

things should be thought of, I believe; it seems more appropriate to say that the *semantics* assigns characters, functions from (reference-) contexts to propositions, to *types* of sentences, while the *token* of a sentence determines the features of the actual (reference-) context needed to determine the actual proposition expressed.

[24] The term is due to Segerberg (1973).

$$\dagger[\underline{A}] \quad = \quad \{\langle w, w'\rangle |\ w' \in K(w)\ \&\ w'[\underline{A}]w'\}$$

The *dagger* is a two-dimensional operator which projects the diagonal of the relation $[\underline{A}]$ into the horizontal (rows are context-worlds, while columns are index-worlds):

$[\underline{A}] =$

	u	v	w
u	1	0	1
v	1	0	1
w	0	1	0

$\dagger\,[\underline{A}] =$

	u	v	w
u	1	0	0
v	1	0	0
w	1	0	0

The application of $\dagger[\underline{A}]$ to the actual world determines the proposition that is true in w' for any w' in $K(w)$ iff $\underline{A}$ uttered at w' is true at w'. In other words, $\dagger[\underline{A}](w)$ is what Stalnaker (1978) calls the *diagonal proposition* expressed by $\underline{A}$ in w. Note that if we assume that for each world w' and w'' in $K(w)$ it holds that $K(w') = K(w'')$,[25] it will be the case that $\dagger[\underline{A}](w') = \dagger[\underline{A}](w'') = \{w' \in K(w) :\ w'[\underline{A}]w'\}$.

With another diagonal operator, '@', we can express that A is *actually* the case:

$$@[\underline{A}] \quad = \quad \{\langle w, w'\rangle |\ w' \in K(w)\ \&\ w[\underline{A}]w\}$$

Note that if $w[\underline{A}]w$, $@[\underline{A}](w) = K(w)$. This operator is normally called the *dthat*, the *upside down dagger*, or simply the *actuality operator*.[26] The *dthat* is a two-dimensional operator that projects the diagonal of the relation $[\underline{A}]$ into the vertical:

$[\underline{A}] =$

	u	v	w
u	1	0	1
v	1	0	1
w	0	1	0

$@[\underline{A}] =$

	u	v	w
u	1	1	1
v	0	0	0
w	0	0	0

It should be obvious that we can define the *dagger* and the *dthat* operators not only on propositional concepts, but also on other two-dimensional intensions (see Appendix A). Equally obvious is it that just as we can distinguish three kinds of rigidity and three kinds of necessity in a Kaplanian framework, there are also several ways in which a token of a term can be called rigid, and a token of a sentence can be called necessary. For instance, the sentence token $\underline{A}$ of A might be called *a priori* necessarily true with respect to context K iff K is a subset of the diagonal proposition expressed by A with respect to K, i.e. if for any w' in K it holds that $\underline{A}$ uttered at w' is true at w'. A final point worth observing is that we have not yet made full use of the assumption that what is *presupposed* by the speaker should be thought of as being part of the context, a fact about the world.[27] There

[25] This is the assumption that presupposition states are *introspective*, and will be defended in chapters 2 and 4.

[26] In chapter 4 I will also introduce *indexed* actuality operators.

[27] This extension of context theory was proposed by Stalnaker (1970b). Note that for the Stalnakerian analysis of diagonalisation, this is indeed an essential part of the context.

will be many occasions in this book, however, where we allow ourselves to come back to this point.

1.9 Solving problems by diagonalisation

We have already seen that distinguishing the roles of contexts and indices makes it possible to explain the distinction in kinds of necessity for true identity statements like *Hesperus is Phosphorus* and *I am here, now*. Let us abbreviate metaphysical necessity by *necessity*, and a priori necessity by *a priori*. If a proposition is not necessarily true or false, it is *contingent*. The first kind of statement is not *a priori*, but if true, it is necessarily true; while the second kind is *a priori* true, but contingent. So, *I am here now* is true in every context in which it is uttered,[28] but need not to be true in every index world with respect to a particular context. It follows that *I didn't have to be here, you know* can still be true.

The *a priori* status of the statement *Stick S is one meter long* can be analyzed along the same lines. It is useful first to distinguish, with Kripke (1972), two ways an identity statement like *E is the N* can be used. It can be used (i) to state the identity of meaning (intension) of the two terms, or (ii) to fix the meaning of one term in terms of the meaning of the other. Thus, sometimes the description *the N* in *E is the N* is used to fix the reference of *E*. This is what is going on in a sentence like *One meter is to be the length of S*. The meaning of the name *one meter* is fixed by the reference-fixing use of *the length of S* by the occasion of utterance. In every context in which the meaning of *one meter* is fixed, it will be the length of stick S, although the length of the stick might have been different from what it actually is. Something similar is going on in *Deep throat is the person who was the source of Woodward and Bernstein's Watergate information*. The reference of the name *Deep Throat* is fixed by the fixing-reference use of the description that follows it. The only difference with the foregoing case is that now *we* don't have any specific individual in mind. Whatever the relevant meaning of *Deep Throat* is, it is clear that the counterfactual *If Haldeman had released the information to the reporters, he would have been Deep Throat* is unacceptable because we consider only counterfactual situations in which Deep Throat is the person who *actually* released the information to the reporters (see Evans (1979)).

The next problem we will consider is that of necessary and impossible propositions. How is it possible that a sentence like *I am Robert* can be informative in some conversational contexts? For instance, how can we explain that I can use this sentence as an informative answer to your question *Who are you?* As far as you know, the world might be such that the person you are asking this question to is Robert, in w, or someone else, in w'. Thus, $K(w) = \{w, w'\}$. Now the above sentence, A, uttered by me, Robert, would express a necessarily true proposition, $[\underline{A}](w)$, in w, but it would express a (necessarily) false proposition, $[\underline{A}](w')$, in w'. According to Gricean conversational rules, and Stalnaker's (1978) first assertion condition, every assertion should express a contingent proposition with

[28] Where a context is either a Kaplanian context, or a world containing an utterance token.

respect to what is presupposed by the speaker. The hearer can conclude that the speaker intended to communicate neither $[\underline{A}](w)$ nor $[\underline{A}](w')$. Moreover, on Grice's and Stalnaker's analysis again, it should be clear to the hearer what proposition is expressed by a given sentence. In each possible world of the context compatible with what is presupposed, the same proposition should be expressed. What could this proposition be? In these cases, Stalnaker (1978) suggests, we should look not at the horizontal proposition expressed in world w, but rather at the diagonal proposition, $[\dagger\underline{A}](w)$. The diagonal proposition expressed is the same in every world of $K(w)$ because it abstracts away from the context-world. If $[\underline{A}](w)$ is necessarily true, but $[\underline{A}](w')$ necessarily false, $[\dagger\underline{A}](w)$ will be contingent. Of course, $[\dagger\underline{A}](w)$ can be different from $[\underline{A}](w)$ in several worlds only if $\underline{A}$ determines a non-constant propositional concept.

Obviously, the diagonal proposition expressed by an identity asserted between two expressions treated as rigid designators can be contingent only if these terms do not necessarily have a constant two-dimensional intension. This is clearly the case for indexicals; in different contexts it might be a different person who is speaking. In the same way, with a token analysis of diagonalisation, Kaplan's paradox of direct reference can be explained. The problem is to explain how a very slow utterance of *This* [pointing to Venus in the morning sky] *is identical with that* [pointing to Venus in the evening sky] can be informative. This can be explained by saying that in some worlds consistent with what is presupposed in the conversation, the token of *this* will not refer to the same object as the token of *that*. The result will be that the hearer is informed that the most salient heavenly body in the morning sky is identical with the most salient heavenly body in the evening sky. So, diagonalisation can explain away some paradoxical consequences of the assumption that indexicals and demonstratives are directly referential. Obviously, if it can be assumed that the intension of proper names and natural kind terms are also context-dependent in this way, we could also solve the problem posed by identity statements between two such terms in such a way. But are the intensions of these terms context-dependent in this way?

Of course the intension of a name is context-dependent. A lot of individuals have the same name, and it depends on the conversational context what the most salient individual meant by the use of a name is. What is more interesting to know, though, is whether the meaning of a proper name can also be world-dependent.[29] In one sense it cannot be denied that the meaning of a proper name is world-dependent. It is a contingent fact about our language that Venus was called *Phosphorus*; if the semantic facts about our world were different it might have been the case that, for instance, Mars was called *Hesperus*. But as Frege (1892) stressed, identity statements between proper names need not be about purely semantic facts of our language. So the question is whether we can assume that the meaning of a proper name is world-dependent, but not just because of the fact that objects could have been called differently.

[29]This might be somewhat misleading, because it seems reasonable to assume that who or what is salient in a particular conversation is a fact about the world.

The externalist theory of reference denies that the meaning of a proper name, or any other kind of expression, is world-dependent in the sense that there is (normally) no description associated with the term that determines its extension through fit with reality. On the other hand, the causal information-theoretic account suggests that there is a body of information associated with the expressions we use. It can then be assumed that the reference of the expression is world-dependent not because in different worlds it may be a different object or stuff that best fits this body of information, but because in different worlds it might have been a different object or stuff that is the dominant source of this body of information.

Remember from section 1.6 that according to the information-theoretic account, we refer with the English expression *water* to H_2O in this world because we *normally* use this term to refer to stuff that has certain observable properties, and *normally* it is only H_2O that has those properties in the actual world. But we also saw that these *normality conditions* are *contingent*; they might be different from world to world. On Twin Earth the normality conditions of the actual world do not obtain: there it is normally not H_2O but XYZ that has the relevant observable properties, and is 'responsible' for the use of the term *water* by Twin-Earth (twin-) English speakers. Something similar holds for proper names.

Once it is assumed that the referent of, for instance, a proper name is world dependent, it is clear that by diagonalisation we can normally account for the informativity of an identity statement like *N is M*, where N and M are both names. In a sense, the reason why the meanings of expressions are world-dependent just depends on semantic facts about the words. Still, we can learn something non-linguistic if we are informed that *N is M*, because even if the exact referent of an expression used in a conversation is not clear, we normally *do* have a pretty good idea about what properties the referents of terms being used have. Thus, if we receive the information that the sentence *Hesperus is Phosphorus* is true, we learn not only some facts about the semantics of English, but also some astronomical facts. We learn that the most salient heavenly body seen in the morning sky is identical with the most salient heavenly body seen in the evening sky, because we already believe and presuppose that we are in a world in which the referents of the relevant expressions have those properties.

It may seem that once we assume that the reference of a proper name is world dependent, we can also immediately account for *negative existential statements* containing proper names. But things are not that easy. If the reference of a proper name is world-dependent only because the dominant source of the information associated with our use of a proper name is not clear, we still seem to presuppose that a dominant source of this information does exist. But isn't this exactly what we claim not to be the case with negative existential statements? Perhaps negative existential statements should be seen not as assertions, but as presupposition *denials* instead. But then, denials are normally reactions to earlier utterances in which the opposite is asserted. This leaves us with the

equally difficult question of how we can appropriately assert contingent propositions with *positive* existential sentences. Donnellan (1974) proposed that with a negative existential statement we simply assert that the proper name has no referent. Stalnaker (1978) offers an attractive way to implement this solution: namely, in terms of his diagonalisation strategy. From Donnellan's discussion it seems that negative existential statements involve only the *mention*, rather than the *use*, of proper names. But when proper names occur in negative existential sentences it seems that the hearer has to understand the singular term in the same way as in a normal use of a proper name. If we use the diagonalisation strategy, we don't have to distinguish different uses of proper names to make use of Donnellan's proposal. The normal use of a proper name presupposes an existing individual that is the dominant source of the relevant information associated with the name. Sometimes, however, one might presuppose it to be possible that the actual source is just, for instance, a character in a novel, an object to which the existence predicate does not apply. In this case, the diagonal proposition associated with a sentence like *N does not exist* will be contingent, and seems to be the right candidate for that which is expressed by such a sentence.

1.10 Self-locating beliefs

1.10.1 The problem of self-locating beliefs

In this chapter we have assumed that a belief state should be represented by a set of possible worlds. According to Perry, however, there is a problem for this analysis which is related to self-locating beliefs. That is, the possible world analysis cannot account for certain kinds of *sameness of beliefs* that different agents might have. Consider crazy Heimson (Perry, 1977), who thinks that he is David Hume. Alone in his study, he says to himself, *I wrote the Treatise*. Of course, he did not. So, contrary to the case in which Hume was thinking this thought, Heimson is thinking something false. However, it seems that we can explain some of Heimson's and Hume's behaviour in the same way if they both think *I wrote the Treatise*. How can the possible world analysis account both for the difference of belief and for the fact that some of their actions can be explained in a similar way?

In terms of a Kaplanian (1989) framework, we can do so by modelling a belief state not by a set of possibilities (indices), but rather by a function from contexts to such a set of indices, a *character*. We can explain some of the actions of both Heimson and Hume in a similar way because they have a belief in common. They both stand in the belief relation to the character expressed by the sentence *I am Hume*. Their beliefs differ, however, because the *propositions* expressed by this sentence if said (or thought) by Heimson and Hume are different. Modelling a belief state by characters can also account for *fine-grained ignorances*, a notion which will be explained presently.

Traditionally, it was assumed that possible worlds could be completely determined

by an impersonal description or eternal sentence. Two possible worlds are the same if they are qualitatively the same. However, Lewis (1979a) showed that belief states cannot be represented by sets of possible worlds understood in this way. Such a representation is not fine-grained enough. We should distinguish more possibilities than there are qualitatively different possible worlds:

> Consider the case of the two gods. They inhabit a certain possible world, and they know exactly which world it is. Therefore they know every proposition that is true at their world. Insofar as knowledge is a propositional attitude, they are omniscient. Still I can imagine them to suffer ignorance: neither one knows which of the two he is. They are not exactly alike. One lives on top of the tallest mountain and throws down manna; the other lives on top of the coldest mountain and throws down thunderbolts. Neither one knows whether he lives on the tallest mountain or on the coldest mountain; nor whether he throws manna or thunderbolts. (Lewis, 1979a, pp. 520-521)

Even if two individuals know exactly what qualitative world they live in, they still might lack certain pieces of knowledge. I will say that such agents are ignorant of certain fine-grained pieces of information. But how, then, can the ignorance of the two gods be accounted for? If belief states are represented by characters the problem disappears: the two sentences *I am the god on the tallest mountain* and *I am the god on the coldest mountain* don't express the same character.

Perry's (1977) proposal has been adopted in cognitive psychology. According to the research strategy in cognitive psychology known as *individualism*, psychological explanations of behaviour can and should be given completely in terms of the internal states of agents. They *should* because what causes the behaviour of the agents are these states. This doesn't mean that these internal states don't have content. They have contents, but (given what has been learned from the Twin Earth stories) these contents cannot be the wide contents, the contents of thoughts determined via externalist means. Different believers can believe different propositions by thinking a thought of the same sentence type. Still, two people who are thinking this have something in common. What they have in common, it is proposed, is a function from contexts to propositions — that is, a character. Thus, psychological explanations *can* be given in terms of internal states only, because of the existence of characters. The contents of internal belief states, the *narrow contents*, are modelled by characters; and what the believer believes, if it is embedded in a specific context, the *wide content* of his belief, is just the result of applying the narrow content to the actual context. Let us denote the internal belief states of Oscar and his twin by $[O]$ and $[TO]$, respectively.[30] The (narrow) contents of their internal states are the same, but

[30] An individualist like Fodor assumes that a belief state should be modelled not by a character, but rather by a *set* of characters.

because they live in different environments, c and c', the intensions of their thoughts are not the same, $[O](c) \neq [TO](c')$.[31]

1.10.2 Fine grained possibilities

According to individualists, belief states should be modelled by something like characters, or better, by sets of characters. But this is problematic if we assume that we should represent a belief state by a set of possibilities, as the pragmatic account of intentionality seems to demand. This is given up, however, when belief states are modelled by (sets of) characters. According to Lewis (1979a) and Stalnaker (1981), we don't have to model belief states by characters to account for the fine-grained ignorances that the two gods have. If we use diagonalisation we can still model a belief state by a set of possibilities.[32] According to Lewis, the gods know what world they live in, but lack knowledge about *who* they are, or *where* they are in a world. He concludes that a belief state can no longer be represented by a set of worlds, a proposition. Analyzing self-locating beliefs, according to Lewis, requires a belief state to be represented by a set of agents, a *property*. The believer has a belief about himself, namely that he possesses a certain property. This property can be that he inhabits a certain world, but it can also be that he is a certain individual, or that he is in a certain position in a world.

Lewis can assume that belief states may be represented by sets of individuals because he assumes that individuals can live in only one possible world. If we don't want to commit ourselves to that assumption, we can say that a belief state should be represented by a set of agent-world pairs, or *centered* worlds. If $\langle a, w\rangle$ is such a pair representing an element of the belief state of some individual Lingens, a is the individual that possesses in w all the properties Lingens ascribes to himself in the actual world. According to Lewis, *de dicto* beliefs and beliefs with essential indexicals are always self-attributions or *de se beliefs*. So, as far as Lingens can tell, he might be the individual a in w. The information that the two gods lack is not what world they live in, but *who* they are. Their belief states can be represented by the following set: $\{\langle gt, w\rangle, \langle gc, w\rangle\}$, where gt is the god on the tallest mountain and gc is the god on the coldest mountain. To analyze other cases of essential indexicals in a similar way, we should in general represent belief states, the belief state of John, for example, by a set of quadruples of entities, where such a quadruple, $\langle a, t, p, w\rangle$, consists not only of the individual a John takes himself to be in w, but also of t, the time he thinks of as 'now' in w, and p, the place he takes himself to be in, in w. Because many different n-tuples can contain the same possible world, Lewis' representation of belief states

[31] This raises the question of when, according to individualists, a belief attribution is true. On the perhaps most straightforward reading, Perry (1977) says that the *truth* of a belief attribution depends on the wide content alone, while the *appropriateness* of the belief attribution depends on narrow content also. If this is the right interpretation of Perry, it looks very much like the neo-Russellian analysis of belief attributions proposed by Salmon (1986), which Perry gave up in Crimmins & Perry (1989).

[32] And note that characters, or propositional concepts, are much finer-grained entities than diagonals.

seems to be finer-grained than the pure possible-worlds account allows for.

According to the pragmatic analysis of attitude states, attitude states are holistic in nature. We do not have a belief box, with several belief objects (however they are modelled) in it, and a desire box, with several desire objects in it. Instead, the attitude state of an agent is modelled by a global belief/desire state, where the belief determines the relevant possibilities and the desire orders these (and other) possibilities with respect to their desirability. If the possibilities needed to model certain beliefs are finer-grained than possible worlds, the question arises whether this fine-grainedness is also needed for the analysis of desire, and thus for the analysis of deliberation. Both questions are answered in the affirmative by Lewis. He convincingly argues that for some deliberations it is important that an agent considers more possibilities than the traditional conception of possible worlds would allow for. It can be that the most useful action to take if you are at one place is different from the one that you would take if you were at another.

By means of Lewis' finer-grained representation of belief states, it is also possible to describe what is special about self-locating beliefs. Self-locating beliefs are special in that they crucially involve not only the world of a possibility, but also something else. But how can Lewis account for the fact that we can explain some of Heimson's and Hume's behaviour in the same way when they both believe *I wrote the Treatise*? According to Lewis we can explain their behaviour in a similar way, because we can characterize their belief states in a similar way. But *how* can we do so?

The simplest way would be to follow Lewis and say that both would self-ascribe the same property. Equivalently, as shown by von Stechow (1984), we might also account for sameness of belief in terms of Kaplan's theory of demonstratives.[33] In Kaplan's theory, sentence type A is true in context c iff $c[A]i(c)$ holds. But then we can associate with sentence type A the set of contexts in which it is true. Let's say that $[\star\ A]$ denotes this set.[34] A context in Kaplan's theory is a set of quadruples like $\langle a, t, p, w\rangle$, a possibility of the same kind as those that Lewis uses. A Lewisian belief state can thus be thought of as a set of Kaplanian contexts. According to Lewis' analysis, Hume and Heimson share a belief because both of their belief states are subsets of [$\star$ I wrote the Treatise].

Von Stechow showed that we can use the $\star$-operator to account for the intuition that Heimson and Hume share a belief. But sameness of belief can also be analyzed within the Kaplanian framework by looking at the *diagonal*. In two-dimensional modal logic, diagonalisation makes sense because contexts and indices are supposed to be of the same kind. Until now we have followed Stalnaker (1978) in assuming that if context and index are of the same kind, we have to make use of a token analysis. But to make contexts and indices of the same type, we don't have to give up a type-analysis when we make indices into finer-grained entities. When A is a sentence, we simply assume that context-index pairs (cips) determine both *what* is expressed by A, and the *truth value* of what is expressed

[33]That is, if we assume that the extension of, for instance, proper names is not world-dependent.

[34]The $\star$-operator is the Kaplanian analogue of the actuality operator in two-dimensional modal logic.

by A. Let's abbreviate such a cip, $\langle c, w\rangle$, by e. Thus, the character of A, $[A]$, can be seen as a relation between cips (van Fraassen, 1979). The diagonal expressed by A in e can now be determined just as before, $\dagger[A](e) = \{e' |\ e' \in K(e)\ \&\ e'[A]e'\}$. Let us now assume that a context is something like a triple, $\langle a, t, p\rangle$, and that the index is a world. This means that a cip is really a quadruple. We have seen that we can read Lewis (1979a) as representing a belief state in precisely this way — by a set of such quadruples. What Heimson and Hume have in common is that each of their belief states is a subset of the following set of cips: [† I wrote the Treatise](e).

It's nice to know that essential indexicals do not force us to give up the traditional view that belief states can be modelled by sets of possibilities, where these possibilities need not be as unspecific as possible worlds. However, traditionally it has been assumed that a believer stands in a relation to some informational content, a proposition, and that this notion is individuated in terms of truth conditions only. There are at least three good reasons, I believe, why the object of belief should be thought of in this way. First, consider sleeping O'Leary (Stalnaker, 1981) locked up in the trunk of his car. He wakes up when the town clock tolls, but isn't sure whether it rings three or four times. *I wonder whether it is now three o'clock*, he says to himself. At nine o'clock he is rescued from the trunk of his car. This time he says to himself *I still wonder what time it was <u>then</u>*. What he wonders about at these two times, it seems reasonable to assume, is the same, a proposition; but on Lewis' account of essential indexicals this reasonable assumption cannot be made. Second, it seems natural to assume that the objects of speech acts and the objects of beliefs are of the same kind, and are propositions, rather than properties. As Stalnaker (1981) noticed, only if the objects of speech acts are propositions can we give a straightforward account of the following kind of conversation:

> Heimson, not so sure anymore whether he is Hume, wants to ask the almost omniscient god on the tallest mountain who he is. Finally reaching the top of the tallest mountain he says to the god "I'm confused and don't know who I am," and then asks "Can you tell me? Who am I?" "You're Heimson, the crazy student," replies the god somewhat impolitely.

The proposition expressed by the answer given by the god on the tallest mountain is a direct answer to Heimson's question. Third, and perhaps most important, informational content should be individuated by truth-conditional content to be able to behave identically in causal and counterfactual constructions. We have seen that according to both the pragmatic, and the information-theoretic account of intentionality the dispositional concept of *belief* is explained in terms of counterfactual dependencies. Consequently, the object of belief should be individuated by truth-conditional content, a set of possible worlds.

1.10.3 Stalnaker's solution

How can we account for the extra fine-grainedness needed to analyze self-locating beliefs in terms of possible worlds individuated only by truth conditions? Stalnaker (1981) suggests that this can be done if we give up the assumption that possible worlds, and/or the relations between possible worlds, can be characterised by completely qualitative means. What Lewis' example of the two gods shows is that we need to distinguish more possibilities than worlds that we can distinguish by qualitative means. But then, once we assume with Kripke (1972) that individuals can exist in more than one possible world independent of their (non-essential) properties, we already have to assume more possible worlds than we can qualitatively distinguish. In particular, there is a distinction between the actual world where d is the god on the tallest mountain and a counterfactual world, qualitatively indiscernible from the actual world, where d is the god on the coldest mountain. The ignorance of d can be modelled as a doubt whether he is in what *we* would call the actual world or this counterfactual world. Crucial for Stalnaker's analysis is, first, the observation that agents who have beliefs are inhabitants of the actual world; and second, the assumption that the subject of the attitude exists not only in the actual world, but also in all worlds that help to characterize his belief state.

Of course, Stalnaker's solution is closely related to Lewis'. Suppose for simplicity that we model a belief state by a set of world-agent pairs. Suppose that d is an individual with a doubt. The belief state of d can then be modelled in a Lewisian way by something like $\{\langle a, w\rangle, \langle b, w\rangle\}$. The only qualitative way in which $\langle a, w\rangle$ differs from $\langle b, w\rangle$ is that in the first possibility, a is the god on the tallest mountain; while in the second, b is the god on the coldest mountain. For the Stalnaker solution, suppose that d is the only individual that exists in the actual world, w, and this counterfactual world, w'.[35] The belief state of d can be modelled by $\{w, w'\}$. The only way in which w differs from w' is that in w, d is the god on the tallest mountain, while in w', he is the god on the coldest mountain. Obviously, there is no substantial difference between the two solutions.[36]

But what about the case of indistinguishable identical twins? Aren't we committed, on a reasonable assumption of supervenience, to claim that their belief states should be represented in exactly the same way? If so, this seems to be a real problem for any externalist position.[37] But then, all we have to explain is that under qualitatively identical circumstances the two twins would act in exactly the same way. This explanation is given in terms of an attributed belief/desire system. We have seen earlier that the pragmatic

[35] Or d and d', two 'individuals' related to each other via a primitive counterpart relation.

[36] That is, if it is assumed that individuals can have singular beliefs about themselves only. If this assumption is not made, it is not clear how belief states can explain behaviour if those states are modelled by sets of possible worlds. Although this assumption is compatible with Stalnaker's first way of describing the puzzle, it is not a very natural assumption.

[37] In a very enlightening discussion of *de se* beliefs, Haas Spohn (1994) has argued against Stalnaker's solution on exactly this ground; although it can explain the difference in belief of the two gods, it cannot explain what indistinguishable twins have in common.

account of intentionality need not completely determine the content of one's thought. The pragmatic account cannot distinguish two belief/desire states that predict, or can explain, the same kind of behaviour in the same way. So, we can substitute one object throughout the whole belief/desire state for another, without predicting any difference in behaviour (Stalnaker, 1984, ch. 1). That in this case we substitute the *agents* of the beliefs for one another doesn't seem to make a crucial difference.[38]

To account for the case of the two gods in terms of possible worlds only, we don't need to rely on diagonalisation: the descriptions *the god on the tallest mountain* and *the god on the coldest mountain* are not considered to be rigid designators. Things are different if both terms that flank the identity sign are thought of in this way. Consider the following example from Perry (1977).

Rudolph Lingens is the amnesiac lost in Stanford Library. Lingens knows a lot about himself, but unfortunately he doesn't know that *he* is the amnesiac lost in Stanford Library. That is, before he has found out, he would not assent to the statement *I am the amnesiac lost in Stanford Library.* But after reading a biography about himself, he believes that Rudolph Lingens is the amnesiac lost in Stanford Library. Suppose now that the proper name *Rudolph Lingens* and the indexical *I* are interpreted as rigid designators. It follows that *I am Rudolph Lingens* expresses either a necessarily true or a necessarily false proposition. But how then can we explain in terms of possible worlds only that Lingens is wondering whether this sentence is true or not?

Because the sentence contains two rigid designators, it seems that there are two ways to solve the problem: by giving up the rigidity of either the proper name or the personal pronoun. If we want to describe the situation by giving up the rigidity of the proper name, we can assume with Stalnaker (1981) that Lingens has known all along who he himself is, the same individual in all possible worlds representing his belief state, but did not know who Rudolph Lingens is. Diagonalisation now has the effect that the rigidity of the name *Rudolph Lingens* is given up. Lingens wonders who the source of the body of information is that he associates with the name *Rudolph Lingens.*

Let's be a bit more explicit. There are two relevant situations. First we have the actual world, w, where d, the actual Lingens and the reader of the biography, is also the subject of the biography, and thus the source of the information that he associates with the name *Rudolph Lingens.* Second, we have the counterfactual world w', where d, the actual Lingens and the reader of the biography, is neither the subject of the biography nor the source of the information he associates with the name. His belief state before he learns that he is Lingens can be characterised by $\{w, w'\}$; after he has learned this, however, world w' is eliminated. Kaplan's (1989) change from cool to hot can be explained in the same way.

[38] A bit more formally: Let w' be an element of $K(d, w)$, the Stalnakerian representation of the belief state of d in w. Is there for this world w' a unique agent-world pair $\langle a, v\rangle$ such that $\langle a, v\rangle$ is an element of the Lewisian representation of the belief state of d in w? If qualitative difference is all that counts for the pragmatic analysis of belief, I think the answer is yes. w' corresponds to $\langle a, v\rangle$ as an element of the representation of d's belief state if for all qualitative properties P, $P(d, w')$ iff $P(a, v)$.

I have just described the situation of Lingens in such a way that he is wondering not who he himself is, but only who Lingens is. Is it also possible to describe the situation in such a way that Lingens is wondering who he himself is without giving up the assumption that belief states should be modelled by sets of possible worlds individuated by truth conditions only?

Stalnaker (1981) argued that this can be done, too, but only if a token-reflexive analysis of indexicals is assumed. According to a token analysis, a context is not an agent, time, place, and world tuple as in Kaplan (1989), and also not simply a world, but a world plus a token of an expression. The referent is then the referent of the token of the expression in that world. In particular, the referent of a token of *I* in a world is the utterer of this token in this world. However, in different worlds it might have been a different person who was the utterer of the token (or a counterpart of it). Thus, in different worlds the personal pronoun *I* might have had a different referent. Now consider Lingens again, who says to himself *I am Lingens*. According to the first way of describing the situation, Lingens has known all along who he himself is, but hasn't known who the referent of the name *Lingens* is. According to the second way, it is the name *Lingens* that refers to the same individual in all worlds characterizing Lingens' belief state, but Lingens is not sure who he himself is — that is, who the utterer of the personal pronoun *I* is. The result will be that Lingens will believe the proposition that can be expressed by *the thinker of the thought token of "I am Lingens" is Lingens*.

Still, you might think that this cannot be the whole solution, because each world of a belief state now contains not more information than that a certain individual is the referent of a specific token, and that this individual satisfies a particular predicate. It is still not clear how Kaplan's learning that he is in a world with this property can explain his change from cool to hot. But this argument does not go through if it is assumed that the agent is *introspective*, i.e., if in all worlds compatible with what the agent believes, there exists an individual that has the same beliefs and other attitudes as the agent has in the actual world, and thinks with the same thought tokens.[39] But for each world compatible with what the agent believes, it seems that the question who this individual then is is still appropriate, given that we have now not made the assumption that it's the same individual in all relevant possible worlds. But the answer to this question is really straightforward: it is just the unique individual in this world that has the same beliefs as the agent himself has in the actual world, and is thinking the same thoughts as this agent too. But what if there are in such a world *two* individuals that have the same beliefs, like Perry's Heimson and Hume? In Perry's (1979) story Heimson believes that he is Hume, and Heimson and Hume agree about all the facts about Hume that could be stated in an impersonal way. The answer is that there can be no worlds in which two such individuals exist, given our assumption that worlds also contain information about the reference of

[39] In terms of possible worlds semantics, a belief state is introspective and consistent, if the accessibility relation that helps to characterize his belief state is *serial*, *transitive* and *euclidean*.

tokens of expressions. If both Heimson and Hume think a thought token of the sentence *I am Hume*, they are thinking something different, because the tokens are different. As a result, the propositions they believe are different, and thus their belief states must differ too.[40]

This latter point suggests that although we can explain on Stalnaker's second solution what Lingens can learn, and in what sense Lewis' two gods have different beliefs, the solution cannot account for cases where two indistinguishable identical twins are involved. But again, this problem doesn't seem insuperable: we should compare not their actual belief states, but their belief states if they would give up certain of their beliefs, in particular their beliefs that certain tokens of expressions exist.

Stalnaker has given two kinds of solutions to the problem of how to analyze indexical belief on the assumption that belief states are to be modelled by sets of possible worlds. It is important to see that the two proposed solutions do not correspond to different kinds of situations, but rather are two different ways of describing the single situation of Lingens thinking *I am Lingens*.

1.11 Belief, and de dicto belief attributions

Traditional wisdom has it that the truth value of an attitude attribution does not depend on the *extension* of the embedded sentence. However, given the assumption that a semantic theory maps surface structures of natural language to semantic values in a rigid way, we seem to be forced to accept that the truth value of a *de dicto* belief attribution cannot even depend on the *intension* of the embedded sentence, since this would lead to inconsistent beliefs in terms of which we would not be able to explain agents' behaviour.

1.11.1 Diagonalisation and aboutness

Consider now a situation where *we* know how to distinguish H_2O from XYZ, but Oscar does not. Although the belief attribution *Oscar believes that water is the best drink for quenching thirst* can be used to explain Oscar's H_2O drinking behaviour, Oscar himself would not be able to make a distinction between the actual world, where what is called *water* by ordinary English speakers is H_2O, and the counterfactual Twin Earth, where XYZ is denoted by *water*. In this sense, his thoughts do not seem to be *about* H_2O, although the causal account seems to predict that they are. We have seen that individualists have concluded that what explains behaviour does not depend on something outside of the agent. What explains Oscar's behaviour is a thought internal to Oscar; and what the content of this thought is, the *narrow content*, can be determined without looking at external circumstances.[41] It is then assumed that, by something like the diagonalisation

[40] Granted, I must make an assumption about thought tokens that you might disagree with.

[41] See Fodor (1987). It should be noted that Fodor defends this position only in this book.

strategy, we can determine what this narrow content is. What we have to ask is what Oscar's sentence *Water is the best drink for quenching thirst* would express according to the semantic rules of Oscar's *language of thought* in different possible worlds. The semantic rules of the language of thought assign to all types of expressions functions from worlds to intentions, Kaplanian characters. According to the semantic rules of Oscar's language of thought, the thought token of the word *water* denotes H_2O in the actual world and XYZ in a counterfactual twin-world. If in both of these worlds what is denoted by *water* is the best drink for quenching thirst, in both of these worlds the belief attribution would be true, and that is why his thought is not about H_2O.

But there are two problems with this argument. First, it is not at all clear that narrow content *can* be determined without looking at external circumstances. Second, it is not obvious that we *need* to explain the behaviour of Oscar by abstracting away from external circumstances. The assumption that we can determine narrow content without looking at external circumstances is based on (i) the Fodorian assumption that we can single out thought tokens, and that the types corresponding to these tokens belong to a language of thought that has a particular semantics; and (ii) the assumption that we can determine the specific function from contexts to intensions for each expression of the language of thought without looking 'outside the head'. Yet it is not clear why the first assumption should be true; and, as Stalnaker (1989) has stressed, Fodor (1987) might be correct in claiming that two expressions of the language of thought have the same character if they determine the same truth conditions in every context, but this says nothing about how to determine the specific function associated with an expression in the language of thought. So, even if there is something like a language of thought, it is not at all clear how the narrow content of expressions in this language can be determined by looking only at the internal state of the agent.[42]

Also, it is not obvious that we *have* to explain the behaviour of Oscar by abstracting away from external conditions. The problem is that although Oscar is not able to distinguish Earth from Twin Earth, or H_2O from XYZ, his thoughts are, from an externalist point of view, about H_2O. How can we account for both intuitions if content is determined by causal means? Note that to account for the first intuition, we don't have to assume that a belief state should be represented by a character,[43] a function from contexts (external conditions) to contents. Just like in the case of self-locating beliefs, we can make use of diagonalisation. To apply the diagonalisation strategy (in the Stalnakerian framework), we have to be able to determine a propositional concept, and to determine that we have to ask, for each of the worlds compatible with what Oscar believes, what proposition would have been expressed by a token of the sentence *Water is the best drink for quenching thirst.* The diagonal of the relevant propositional concept will be true in the actual world, where *water*

[42]For more on this, see Stalnaker (1989, 1990b). Note also that the pragmatic, or functional, account of intentionality is essentially externalistic.

[43]Or by a set of characters.

denotes H_2O, but also in a counterfactual world where *water* denotes XYZ, a natural kind that looks exactly like the stuff *we* call *water*. It is this diagonal proposition that seems like a reasonable candidate for representing the psychological content of Oscar's thought from his own point of view.

But by making use of diagonalisation don't we predict that the content of one's thought is *independent* of external conditions, just like individualists required? No we do not! According to the causal information-theoretic account of intentionality, even something like the narrow content of one's belief state, the content of the belief from the believer's point of view, is dependent on facts of the environment. But this does not mean that the diagonalisation strategy cannot be used. Remember from section 1.6 that according to the causal information-theoretic account, someone believes that P means that he is in a certain state that *under normal conditions* he would be in only if P. But we have seen that both conditions external to the agent and conditions related to the internal functioning of the representational mechanism belong to these normal or ideal conditions. The relevant normality condition related to the internal functioning of the representational mechanism is in this case the ability of *the agent* to distinguish H_2O from other relevant liquids.[44] If in determining normality conditions we demand that the facts about the agent be *normal for him*, we are, as it were, evaluating his belief, and the belief attribution, from the agent's point of view, and using the diagonalisation strategy. Still, the content of his thought is dependent on external conditions, and is explained by the causal information-theoretic account.

But if we must determine the normal conditions with respect to Oscar's internal functioning of the representational mechanism, as is normal for him, in order to correctly characterize his belief state, aren't we committed to the claim that Oscar's beliefs are not about H_2O at all? How can we account for the intuition that the belief attribution *Oscar believes that water is the best drink for quenching thirst* is both true and still *about* H_2O? To account for this intuition, we have to remember that according to the information-theoretic account of content, the indication relation is analyzed in terms of nested counterfactual conditionals: *if* conditions are normal, and *if* various alternatives to P were true, then the believer would be in various alternative states. This analysis suggests that it is not only the relevant normality conditions, but also the set of relevant alternatives that are context dependent. According to Dretske (1970, 1981), knowledge and belief attributions are essentially contrastive. The belief attribution *Oscar believes that water is the best drink for quenching thirst* is true if Oscar is able to distinguish those alternatives consistent with the relevant normality conditions where water is the best drink for quenching thirst from those where it is not, and count only the former as true. The set of relevant alternatives

[44]Stalnaker (1984) suggests that there are *two* ways in which the internal functioning of the representational mechanism might be normal: (i) it functions normal if it functions as it usually functions by the *agent*, and (ii) it functions normally if it functions as it functions for most of *us*. For *de dicto* beliefs, and belief attributions, the first way seems to be relevant, while for *de re* beliefs, and belief attributions, it seems to be the second way that counts.

depends on what *we* consider to be normal. In normal situations, only these possibilities are consistent with the relevant normal conditions in which what *we* presuppose about the denotation of *water* holds. Because *we* presuppose that water is H_2O, there will be no relevant alternative considered where water is XYZ; thus Oscar's belief can be said to be *about* H_2O, although he cannot distinguish H_2O form XYZ. Alternative worlds where *water* denotes not H_2O but XYZ are considered only when critical questions about the theory of meaning are considered.

What this suggests is that not only the proposition expressed by the embedded sentence of a belief attribution is context dependent, but that also the set of alternatives that model what the agent believes in the world, and thus the way to represent the facts that determine whether the belief attribution is true or false, is dependent on the conversational context (cf. Stalnaker, 1988).[45]

That the set of relevant alternatives depends on the conversational context is also relevant to the analysis of *knowledge* (see Dretske, 1970, 1981) and to certain cases in which we can attribute to different agents the *same belief* (see Stalnaker, 1984). With respect to the first issue, we must be able to account for the intuition that some knowledge attributions are true; but we must also be able to address certain sceptical doubts. This can be done as follows: A knowledge attribution can be true, because we normally consider only possibilities consistent with what *we* presuppose to be normal. Sometimes, however, one of these presuppositions, or channel conditions, is called into question. Once this is done — and this is typically done by a sceptic — more possibilities will become relevant. In such cases we ask more of the agent by presupposing less. The agent must have finer-grained discriminating capacities for the knowledge attribution to be true than he has needed before the relevant normality condition is called into question.[46] To account for sameness of belief, it is crucial, I think, that the relevant alternative possibilities consistent with the normal conditions are only as fine-grained as the conversational context asks for. If in discussing what the agent believes only a few issues are relevant, we don't have to distinguish a lot of alternative states of the world. Suppose that the issue of a discussion is whether two individuals have a belief in common. Suppose also that one agent, a, has a more complex representational mechanism than the other agent, b, has. Agent a tends to be in different internal states when P is the case from when Q is the case, while agent b does not. Suppose now that agent a is in a state that carries the information that P, and agent b is in a state that carries the information that P or Q. If we assume that two states' containing the same information is a necessary condition for individuals in those respective states to have the same beliefs, then it seems that there is a difference in belief. Still, in a context where the difference between P and Q is not relevant, we might say that a and b have the same beliefs. What we do in those situations is to make the set of relevant

[45]This suggests that the latter kind of context-dependence should be accounted for in a similar way as we should account for vagueness.

[46]See also chapter 6, section 2.3.

possibilities by which we have represented the belief state of a as coarse-grained as the set of possibilities used to characterize b's belief state. If we do so, we can say that both beliefs are identical. In this way we can sometimes appropriately say that a human being and a dog have the same beliefs, though different discriminating capacities; and we can also explain why we can sometimes truthfully attribute the same beliefs to two individuals about a certain topic, although one individual is an expert in the field and the other is not.

1.11.2 Diagonalisation and partly linguistic beliefs

In the case of Putnam's Twin Earth example, the question was how we could account for the intuition that the belief attribution *Oscar believes that water is the best drink for quenching thirst* is both true and still *about* H_2O. To account for the truth of the belief attribution we assumed that we look at the *diagonal* proposition expressed by the embedded sentence. To account for the intuition that his belief is *about* H_2O, we needed to determine the relevant normality conditions that are normal for *us*, that is, we had to assume that the only worlds that are relevant to the conversational context are those in which 'water' denotes H_2O. Sometimes, however, we don't want to consider only worlds where a term only denotes a particular individual or stuff. Consider Burge's (1979) Bert, an English speaker who has arthritis. Unfortunately, Bert does not know that arthritis is, by definition, a disease of the joints. He says, and apparently believes, that he has recently developed arthritis in his thigh, which is impossible. Still, when we know that arthritis is a disease of the joints only, the belief attribution *Bert believes that he has arthritis in his thigh* seems to express a contingently true proposition. How can we account for this from an externalist point of view? Bert doesn't believe that he has what is, by definition, a disease of the joints in his thigh. Also, his belief is intuitively not *about* arthritis. But if his belief is not about arthritis, how can the facts about Bert be consistent with an externalist account of content? But this is not a real problem. True, his thoughts are not about arthritis, but that doesn't mean that his thoughts do not depend on external conditions. His thoughts are about a more general disease, a disease one can also have in one's thigh. Nice, but how can an externalist account for the truth of the above belief attribution? Simply by not limiting the relevant alternatives to worlds where the word 'arthritis' really denotes arthritis, a disease one can only have in one's joints. Setting the normality conditions as normal for us will make the belief attribution trivially false, so by Gricean reasoning we conclude that this is not the way we should proceed. We saw above that by not limiting the relevant alternatives to the ones compatible with what *we* believe about the denotation of 'arthritis', we set the normal conditions related to the internal functioning of the agent's representational mechanism. In this case there is something wrong with Bert's use of English. In the dialect of English that he speaks, *arthritis* does not denote a disease of the joints only, but a more general disease. So, if we attribute to Bert the belief that he has arthritis in his thigh, we must not determine a propositional concept with respect to the worlds in which the normal conditions with respect to English

in our world hold, but with respect to a slightly bigger set of worlds, in which the normal conditions with respect to the dialect of English that Bert seems to speak hold.[47] For the belief attribution to be true, it must be the case that in the actual world the set of worlds consistent with what Bert believes is a subset of the diagonal proposition of the propositional concept determined above. Just as in the case for Oscar, it also holds here that more belief attributions can be true if we presuppose less about the relevant normal conditions.

Of course, the diagonalisation strategy might also be used in case the attribution is purely linguistic. To determine the propositional concept expressed by the embedded sentence in the belief attribution *John believes that a fortnight is a period of ten days*, I am not determining for each world in the relevant context what the source of the information is that we associate with the term *fortnight*. The only thing that seems to count in these cases is the description of what is called *a fortnight* by ordinary English speakers. So, the belief attribution is true iff John believes that what ordinary English speakers call *a fortnight* is a period of ten days. We can conclude that we can attribute a belief about linguistic practice to agents without explicitly using metalinguistic terms. Of course, if the description that counts is a description about the use of a term in a certain language, it is to be expected that we cannot always translate the sentence by which the belief attribution is made into another language, and expect that we could attribute the same belief to the agent with this other sentence as we can with the original one. For instance, as noted by Church (1954), we cannot attribute the same belief by means of a German translation of the above sentence as we can by means of the original sentence. Because the translation of *a fortnight* into German is the same as the translation of *a period of fourteen days*, nobody would (*de dicto*) believe what would be attributed by a German translation of the original sentence. We can conclude that what is attributed in a belief attribution might crucially depend on the language used in the attribution. The reason is that although beliefs are always about content, sometimes this content might be about form.

Stalnaker (1972, 1984, 1990a) suggests that the problem of *mathematical belief* can also be partly solved by the diagonalisation strategy. Possible worlds semantics is committed to the view that there are only two mathematical propositions, one that is necessarily true and one that is necessarily false. But in this way there can be no doubt or error about mathematics: we can doubt only contingent propositions. Some have concluded that this problem shows that belief states cannot be modelled by sets of possible worlds: we must take into account not only *what* is believed, but also *how* the agent represents what he believes. A belief attribution is true if both the content and the form of the embedded clause match the belief state of the agent. But this seems to be the wrong way to think of things; beliefs and doubts are always *about* something, and it is only content that counts. What can this doubt or false belief be *about*? The first thing to note is that fine-grained distinctions between logical truths make a difference only to language-using intentional

[47] See also van Fraassen (1979) and Haas Spohn (1994).

systems. If what is expressed by P and what is expressed by Q are necessarily equivalent, although the agent believes the one but not the other, it seems that the agent doesn't have the information necessary to see that these two clauses express equivalent propositions. What this suggests is that the agent's beliefs and doubts are not about what would be expressed by P and Q in w, the actual world, but about the *semantic information* necessary to determine what they express. Let A be a logical statement the truth of which agent x doubts. Stalnaker suggests that x's doubt is not related to the *proposition* that the logical statement actually expresses, but about the relation between *statements* and what they express, the semantic information. If you are in doubt about a mathematical statement, you doubt whether the statement expresses the necessary proposition. The diagonal proposition mirrors this, because in other worlds the semantic rules might have been different. An agent can be in doubt about a mathematical statement if in one of the worlds representing his belief state, the words used in the statement mean something different from what they actually mean.

Although it doesn't seem unreasonable to assume that mathematics is about semantic structure, surely mathematics cannot be just about the specific ways in which mathematical statements are expressed. If Ralph and Pierre say to themselves respectively *Seven plus five equals twelve* and *Sept plus cinq fait douze* they intuitively have the same mathematical belief. What can this object of belief be, if it is not a necessarily true proposition? Stalnaker (1972, 1990a) suggests that mathematics is not so much about the relation between particular tokens of sentences and the proposition they express, as about the relation between more abstract structures that some but not all mathematical statements share and the proposition they express. On a certain level of abstraction, the above English and French sentences share the same structure, and what Ralph and Pierre have in common is that they both believe that sentences that have this structure express the necessary proposition. This suggestion can be analyzed in terms of the diagonalisation strategy, because this strategy accounts for beliefs about the relation between certain representations and the contents of these representations; and these representations need not be particular linguistic entities but can be more abstract representations too.[48] Of course,

[48]Bäuerle & Cresswell (1984) have argued that the diagonalisation strategy cannot solve the problem of mathematical belief because "It seems very implausible to suppose that when someone mistakenly believes that $14 + 23 = 47$ the belief world of that person is a world in which this expression has a different meaning and expresses a truth. For one may well believe *that* without believing that $14 + 23 = 47$. If one believes that the sentence 'Figs fly' is true because one believes that 'pigs' is the word for birds it cannot be concluded that one believes that pigs fly." I agree that in general belief attributions are not about the relation between sentential tokens of the embedded sentences and their semantic values. But this doesn't mean that they never are, nor that they are sometimes about the relations between sets of tokens that share a certain structure and their semantic values. Bäuerle & Cresswell are suggesting (following Cresswell & von Stechow (1982)) that belief attributions like *John believes that 14 + 23 = 47* are about the actual numbers 14, 23 and 47, and not so much about the language. I am not sure what it means to have *de re* beliefs about numbers, but if it means having beliefs about the structures shared by certain tokens of expressions on a certain level of abstraction, then the two solutions might well come down to the

this suggestion by itself will not solve the problem of equivalence and deduction posed by the possible world framework:

> It will not save us from mathematical omniscience to any interesting degree. Given a formal system, its axiom wffs, and its rules of wff-formation and derivation, the theoremhood or nontheoremhood of given wffs follows logically. Thus if I am logically omniscient, know the axiom sentences and rules of derivation and sentence formation of a given mathematical system, and if I am given a theorem sentence, I will, as soon as I identify the sentences in question, know that it is a derivable theorem sentence. (Powers, 1976, p. 100)

But we have seen above that this problem might be partially solved if we assume with Stalnaker (1984) that deduction is the process of integrating different compartments of one's belief state.[49]

1.11.3 Diagonalisation and proper names

The above use of the diagonalisation strategy can also be used for belief attributions involving proper names (Stalnaker, 1987). Suppose that N and M are two proper names that in the actual world refer to different objects. Suppose also that *we* associate the body of information D with the name M, and that John also associates this with this name. Suppose now that I say *John believes that N is M*. To determine the propositional concept expressed by the embedded sentence, we have to ask, for each of the relevant worlds, what would be asserted by *N is M* if it were uttered in this world. The relevant worlds will typically be the worlds compatible with what *we* presuppose John believes.[50] Let us assume that N is *Mars*, M is *Hesperus*, and D is the information that corresponds to the way *we* and John are acquainted with Venus as seen in the evening sky. In some of the relevant worlds, it is not Venus but Mars that is the source of this information. In this case, the diagonal proposition expressed by the embedded sentence in its context of interpretation will be contingent. The belief attribution was appropriate, because this diagonal proposition is true in some but not all of the worlds in the context of interpretation for the embedded sentence. The belief attribution itself is true in those worlds in which the set of worlds that characterize the belief state of John in that world is a subset of the relevant diagonal proposition expressed by *N is M*. If the belief attribution is true, John believes that Mars is the most salient heavenly body seen in the evening sky.

---

same.

[49] A complementary strategy would be to distinguish between tacit and active beliefs, and say that someone actively believes that A if he tacitly believes A and if A is one of the propositions that he is aware of. See section 1.5 for more motivation, and Fagin & Halpern (1988) and Thijsse (1992) for formal accounts.

[50] cf. Stalnaker (1988).

It is sometimes assumed that the diagonalisation strategy can account for belief attributions only when the subject matter of the belief attributed is linguistic in kind. True, the diagonalisation strategy can account for belief attributions only when what is at issue is the relation between a certain representation and its content. But then, not all representations are linguistic representations; thought tokens are representations, too. For this reason it doesn't matter whether or not the agent that the attribution is about speaks the same language as I, the attributer, do. In the example discussed above, for instance, all that counts is that it is presupposed that John is *acquainted* with Venus in the same way as *we* are acquainted with the source of the body of information that *we* associate with the term *Hesperus*. In this way we can account for the intuition that we have attributed to John a belief in an astronomical fact.

Diagonalisation is a useful strategy to account for *de dicto* belief attributions where the wide content of the embedded sentence seems to result in a notion of belief that is too specific or not specific enough to explain the agent's behaviour appropriately. Until now we have used diagonalisation for the analysis of belief attributions with, for instance, singular terms, when the agent believes that the relevant term has a *unique* dominant source. However, this presupposition cannot always be made for the analysis of *de dicto* belief attributions with singular terms involved.

Consider Kripke's case of Pierre again. The problem was that the names *Londres* in French and *London* in English seems to have the same meaning — not only the same extension, but also the same intension. But then, how can we escape the conclusion that Pierre has inconsistent beliefs if he is inclined to say both *Londres est jolie* and *London is ugly*? According to the diagonalisation strategy that we have been using, the answer seems straightforward. In the worlds consistent with what Pierre believes as far as we presuppose, the names *Londres* and *London* denote different cities. Moreover, in these worlds the city called *Londres* is beautiful, but the city called *London* is ugly. That's why, according to this strategy, (4a) and (4b) are true:

(4) a. Pierre croit que Londres est jolie.

b. Pierre believes that London is ugly.

In this case, it is natural to assume that the intensions associated with the names *Londres* and *London* by normal members of the French and English linguistic communities, respectively, are the same. It follows that the diagonalisation solution to the puzzle assumes that beliefs are at least partly linguistic. From Church's (1954) discussion, we have seen that the diagonalisation strategy can account for linguistic beliefs without being stated in explicit metalinguistic terms. But we also saw that there was something special about such cases. We cannot always translate the sentence by which the attribution is made into another language and make the same attribution with this translated sentence; our ability to do so depends on the conversational context. Such a translation is not allowed, for instance, if it is known that the agent associates a relevantly different body of information

with some of the terms used. It seems that in this way we can account for the fact that given that we know the facts that Kripke has given us, we cannot in this context infer from the proposition given in (4a) that given in (5), and derive a contradiction together with (4b):[51]

(5) Pierre believes that London is beautiful.

But we should be cautious here. In the example discussed by Church, translation is not allowed because the attributed belief is completely about the language. In Kripke's case of Pierre, however, this seems not to be the case. The two relevant beliefs that Pierre has are both *about* London; he is just acquainted with London in two different ways, and associates with those two acquaintance relations two different names. But if his beliefs can be said to be really *about* London, it cannot be claimed that translation does not preserve truth value. The best that can be said is that (5) is not a very natural way to state Pierre's belief in the given conversational context.

I believe that this is indeed one justifiable way to react to the puzzle.[52] However, it is not in accordance with Kripke's explicit claim that he was only concerned with *de dicto* belief attributions. I want to propose that if we really want to analyze the sentences (4a) and (4b) as *de dicto* attributions such that we don't attribute to Pierre inconsistent beliefs, we have to use the diagonalisation solution. But how can we do this if the agent associates with the word *two* individuals, while we do not?

To focus the discussion, let's consider the case of Kripke's (1979) Peter. Peter has heard of a great musician named Paderewski. So he is inclined to say *Paderewski is a great musician*. We can conclude

(6) Peter believes that Paderewski is a great musician.

In a different conversational context we learn that Peter has heard of a politician with the name *Paderewski*. We know that he thinks that all politicians are bad musicians, and why should this one be an exception? We can thus also conclude

(7) Peter believes that Paderewski is a bad musician.

In fact, however, the politician named *Paderewski* and the famous musician are the same person. Because Peter would not associate a single individual with the name *Paderewski* in the worlds that might be the actual world as far as he can tell, it seems that the diagonalisation strategy cannot be used to account for this example. But once we assume that the *language* that Peter speaks is different from ours, once we make use of a *token analysis*, the diagonalisation strategy might still be used.

[51] Muskens (1989) argues that we should blame the translation principle for this puzzle.

[52] See also Lewis (1981) and Lerner & Zimmermann (1984).

First, let us assume that we make use of the extended Kaplanian framework of van Fraassen (1979), and Haas Spohn (1994). In that case, following Haas Spohn, we might say that a belief state is represented by a set of possibilities, each containing a 'language'. It is only reasonable to assume that all these possibilities contain the same 'language', perhaps the agent's language of thought. If we now interpret the belief attributions (6) and (7), we look for each $\langle a, t, p, l, w\rangle$ in the set representing Peter's belief state what would be expressed by their embedded sentences. The first thing to note is that on this analysis, we have to assume that when the speaker is uttering the embedded sentences, he is not necessarily speaking English; whether he speaks English or not is, according to the proposal at hand, not at all relevant for evaluating the embedded clauses, and thus for evaluating the whole belief attributions. This remarkable feature is behind all solutions where diagonalisation is used to account for the non-rigidity of proper names and common nouns within a Kaplanian framework. But if we want to account for the (internal) consistency of Peter's beliefs, we have to make two extra assumptions. First, we have to assume that in distinction with ordinary English, Peter's language of thought contains two terms 'Paderewski', *Paderewski*$_1$ and *Paderewski*$_2$. Second, we have to assume that when we assert (6) in the situation sketched by Kripke, *we* have a different term in mind than when we assert (7). Obviously, according to this solution, the translation principle is responsible for the puzzle.

According to a *token analysis*, we don't have to make the speculative assumptions that Peter has a language of thought with a specific semantics, and that to evaluate the embedded sentences only Peter's language of thought was relevant. But how, then, can we account for the (internal) consistency of Peter's thoughts? This can be done by assuming that although *the speaker* associates with the name *Paderewski* both the information that its bearer is a musician, and the information that it is a politician of which Peter has heard, he assumes that in the different belief attributions, different pieces of information are relevant.

Whether we make use of diagonalisation within a Kaplanian framework to solve the puzzle, as Haas Spohn does, or by making use of a token analysis, in both cases we assume that Peter has two representations of Paderewski, and that only one of the two representations is relevant. But how do *we* as hearers determine which of the two representations is relevant in which of the belief attributions? It seems only reasonable to follow van Fraassen (1979), and assume that this is dependent on the conversational situation; which one of Peter's representations of Paderewski is most salient in the relevant conversational situation. After all, Kripke's case of Pierre, for example, only really seems puzzling when we have talked about Pierre *both* as someone who grew up monolingually in Paris, *and* as someone who settles in London. Had we only given one half of the story, it would have been clear to the participants of the conversation which of Pierre's representations of London the speaker meant.

As a result, even if for the analysis of a *de dicto* belief attribution with a singular

term involved it is not the case that the agent believes that the relevant term has only a *unique* dominant source, we can still make use of diagonalisation. The reason is that what counts is not so much that *the agent* actually believes that the term has a unique dominant source, but rather that it is *presupposed* in the relevant conversation that the agent believes that the term has a unique dominant source, or even that there is a unique most *salient* representation of which it is presupposed that the agent associates it with the term. As a result, we can account for the intuition that in different conversational situations we can communicate with the same belief attribution different propositions, although the agent has not changed his mind.[53]

We have seen that the causal information-theoretic account of content gives rise to two different kinds of problems. In the Twin Earth stories the predicted wide content is for some purposes individuated *too specifically*; while for the cases that we have just discussed the predicted wide content is *not* individuated *specifically enough*. The diagonalisation strategy seems to be very useful for the former cases, and it can also successfully account for a lot of cases of the latter kind. Unfortunately, as we will see in the next section, it cannot account for all of these cases. But in those cases where diagonalisation doesn't help, how should we analyze belief attributions where the wide content is too coarse-grained? The problem is an old one: it is the problem of *de re* belief attributions, belief attributions made about particular objects.

1.12 De re belief attributions

1.12.1 Quine's problem

Quine (1953) claimed that modal statements cannot be made about particular individuals, because that would give rise to *Aristotelian essentialism*; the view that an object may have some of its non-trivial properties by necessity, independently of how the object is described.[54] For that reason, Quine argues, we should not make use of quantified modal logic, for that attributes properties to individuals by necessity. But quantified modal logic really gives rise to *two* questions: The *first* question is whether we should be able to *interpret statements* that attribute properties to particular objects by necessity, and if so, *how*? The *second* question is whether some individuals actually *do* have some of their properties by necessity.

To *make sense* of the question whether object d has property P by necessity or not, we have to decide whether this means that in all relevant worlds we have to look at whether the object d itself has property P in that world, or whether we have to look at a/the *counterpart* of d in that world. Even if we don't assume that objects can inhabit more than one world, we still have to decide (i) whether an individual can have more than one

[53] See also Stalnaker (1988), and section 1.13 for a similar point about *de re* belief attributions.

[54] Of course, there will be trivial properties that objects have by necessity, such as being self-identical.

counterpart in a world, or not, and (ii) *how* this counterparthood should be understood.

It seems obvious that when we take the *haecceitistic* view that there is something about an individual that underlies its (qualitative) properties, we already presuppose a positive answer to the second question, the question whether some individuals actually *do* have some of their properties by necessity.[55] The important thing to note, however, is that the question also makes sense for proponents of a counterpart theory; there is no principled reason why counterparthood should be understood in terms of qualitative similarity.[56]

What is perhaps more important to note is that counterpart theory is perfectly compatible with the causal-information theoretic analysis of aboutness. At first sight it might seem as if Kripke's (1972) arguments against counterpart theory follow immediately from his *causal theory* of reference; the causal theory suggested that proper names should be treated as *rigid designators*. But the view that names should be treated as rigid designators does *not follow* from the causal theory of reference. The causal theory only says that proper names refer in a world to their causal origin *in that world*, and does not exclude the possibility that modal statements about the referent of a proper name should be interpreted by means of counterpart relations.

My prime interest in this section is not whether *objects* can have properties by *necessity*, nor how we should interpret *statements* attributing properties to particular objects by necessity. Still, the same issues come up for belief and belief attributions: Can agents have beliefs *about* particular individuals, and how should we interpret *statements* attributing beliefs about particular individuals to agents? In section 1.4 - 1.6 we have already discussed how to make sense of the intuition that agents can have information, or beliefs, *about* particular individuals. In this section I will discuss how to interpret belief *attributions* about particular individuals, *de re* belief attributions.

Whereas Quine thought that modal statements about particular objects are not possible, he admitted that *belief attributions* about particular objects can be made. But he also showed that the most obvious way this can be accounted for within intensional logic would lead to inconsistencies. And indeed, the assumption that for a belief attribution that is about a particular individual, this individual is referred to in all worlds compatible with what the agent believes seems to lead to embarrassing results.

Consider Quine's (1956) Ralph who, one evening, sees a man with a brown hat whose suspicious behaviour leads Ralph to believe that the man is a spy. On another occasion, Ralph sees the same man at the beach, but he does not recognize him as the same man; and the thought that the man he sees at the beach is a spy does not even occur to him. Intuitively, we can attribute his beliefs by saying (8) and (9):

(8) Ralph believes of the man with the brown hat that he is a spy

[55] Of course, Kripke (1972) argued in favor of haecceitism, and made sense of this by assuming that individuals can inhabit more than one possible world, and he also gave some particular examples of properties individuals will have by necessity.

[56] Cf. Stalnaker (1986).

(9) Ralph doesn't believe of the man he saw at the beach that he is a spy

But now the story goes on. In fact, the man with the hat who is later seen at the beach happens to be Ortcutt. So we seem to be allowed to infer (10) from (8), and (11) from (9):

(10) Ralph believes of Ortcutt that he is a spy.

(11) Ralph doesn't believe of Ortcutt that he is a spy.

Now, does Ralph believe that Ortcutt is a spy or not? Or better, how can we account for the beliefs attributed to Ralph that seem to be about Ortcutt without concluding that Ralph is irrational?

The example of Ralph is very similar to Kripke's examples of Pierre and Peter. In the latter case it could be argued that diagonalisation can solve the problem. However, it is not clear how diagonalisation could help to account for Quine's example; Ralph has never heard the name *Ortcutt*, and so he doesn't associate any individual with the name in his belief worlds, and we definitely want to claim that Ralph's beliefs are *about* Ortcutt.[57]

A natural reply to Quine's Ortcutt problem would be to demand that a *de re* attribution can be truly made only if the agent *knows* the object that the belief is about. But then the question arises what it means to have this kind of knowledge. Perhaps we should follow Russell (1905) who claimed that *every proposition which we can understand must be composed wholly of constituents with which we are acquainted*. This would suggest that one can have a belief *about* an individual only if one is *acquainted* with that individual. But acquaintance by itself does not solve the Ortcutt problem. It is reasonable to assume that Ralph is acquainted with Ortcutt. The problem is that he is acquainted with Ortcutt in *two different ways*, and that he doesn't know that a single individual is the source of those two relevant bodies of information. To solve the Ortcutt problem in a purely Russellian framework, we would have to assume that an agent is acquainted with an object only if

[57]This point can and has been made for all kinds of terms that are treated by Kripke and others as rigid designators. For demonstratives, consider the Esa Saarinen example:

> Esa Saarinen, on a semester's visit to the Philosophy Department of UCLA, has told his wife that he is going for the next two days to San Diego. As a matter of fact he is partaking in a punk show in downtown L.A., to which he has invited several of his philosophical friends. Unexpectedly Esa's wife has come to see the show too. But she doesn't recognize the heavily transformed Esa as her husband, when he appears on stage and continues to think that he is in San Diego. Kaplan, sitting closely behind her in the audience can then whisper to his neighbour, pointing first at her and then at the person on the stage, "She believes that he/that man is in San Diego." (Kamp, lecture notes)

Just as we did with Quine's case of Ralph, we can ask: Does or doesn't Esa Saarinen's wife believe that he is in San Diego? Just as in Quine's case, no simple answer *yes* or *no* seems appropriate, and diagonalisation cannot be used: Esa's wife does not hear Kaplan's use of *he/ that man*, nor does she see his pointing, and so does not associate any particular body of information with Kaplan's use of the demonstrative. Still, we want to say that Esa's wife has beliefs *about* Esa.

such cases of mistaken identity are impossible. This seems to be what Russell had in mind; a subject can be acquainted only with objects with which he is in sensory contact. In the strategy proposed by Russell, believers stand in relations with the *content* of the embedded sentence — that is, propositions. A *de re* belief attribution is true only if the agent stands in the belief relation to a proposition about a particular object. Such propositions about particular objects are known as *Russellian* or *singular propositions*. An agent can grasp such a proposition, only if he is acquainted with the object that the proposition is about, where acquaintance means that mistaken identity is impossible. We might say, then, that a *de re* belief attribution is *false* if the agent to which the belief is attributed is not acquainted in this strong way with the object that the belief attribution is about.

Note that given Quine's story, this Russellian account predicts that neither (10) nor (11) will be true. Both of them will be false because Ralph doesn't know that a single individual, Ortcutt, is the source of the two relevant bodies of information; he knows the identity of neither the man with the brown hat nor the man seen at the beach. According to this picture, *de re* attributions can be truly made only if the possibility of mistaken identity does not exist. But that condition is very hard to satisfy. The suggestion that being in sensory contact with an object is enough to make mistaken identity impossible is simply wrong. Consider the following example from Evans:

> Suppose a person can see two views of what is in fact one very long ship, through two windows in the room in which he is sitting. He may be prepared to accept 'That ship was built in Japan' (pointing through one window), but not prepared to accept 'That ship was built in Japan' (pointing through the other window). Now suppose we try to describe this situation in terms of the ordered-couple conception of Russellian thought. We have a single proposition or thought-content — ⟨the ship in question, the property of having been built in Japan⟩ — to which the subject both has and fails to have the relation corresponding to the notion of belief. Not only does this fail to give any intelligible characterization of the subject's state of mind; it appears to be actually contradictory. By constructing cases of this kind, it is not difficult to argue, given the assumption that Russellian thoughts must be representable in the ordered-couple way, that there is very little applicability, and perhaps no applicability at all, for the notion of Russellian thoughts outside Russell's own narrow limits. (Evans, 1982, p. 84)

Russellians must conclude that there are almost no true *de re* belief attributions. This conclusion, however, seems to be false. Suppose we tell only one half of the story. One evening, Ralph sees a man with a brown hat who behaves suspiciously and who, he comes to believe, is a spy. Ralph has never heard the name *Ortcutt*, but in fact it is the person named *Ortcutt* who is the suspiciously-behaving man with the brown hat whom Ralph has seen earlier. In these circumstances the following belief attributions seem to be appropriate and true:

(12) a. Ralph believes of the man with the brown hat that he is a spy.

b. Ralph believes of Ortcutt that he is a spy.

From both ascriptions we can conclude

(13) There is someone of whom Ralph believes that he is a spy.

In this case, even if Ralph has no discriminating knowledge about Ortcutt, the above *de re* belief attributions still seem to be appropriate.

This, then, raises the question how we should account for the fact that Ralph might have two beliefs *about* the same individual that are (apparently) mutually inconsistent.

One reaction would be to assume that Ralph really believes propositions that are mutually inconsistent, and thus that his belief state itself is internally inconsistent. As we have seen in sections 1.5 and 1.12, Stalnaker (1984) and others have argued that the beliefs of an agent should be modelled by a fragmented *cluster* of belief states. These clusters are also used to account for the fact that agents might have *mutually inconsistent* beliefs. Each of these states themselves are modelled by sets of possible worlds and are thus consistent, but two different elements of a cluster might be inconsistent with each other. Now it might be argued that Quine's puzzle of Ortcutt is just a special case of this inconsistency that believers show, and thus should be accounted for in terms of this cluster model too.

Although this strategy seems attractive, because straightforwardly in line with the Kripkean view that proper names should be treated as rigid designators, I won't adopt it. First, because I don't think that the best way to account for Ralph's wondering whether the man he saw at the beach is the same man as the man with the brown hat is done in terms of fragmented belief states. Second, because similar problems arise to account for metaphysical necessity that, arguably, cannot be solved in this way.[58] Third, and most important for me, adopting this account would make it difficult to make a connection between the analysis of *de re* belief attributions and modern theories of discourse, to be discussed in the next chapter, which, according to a realistic conception of these theories, can be thought of as having different representations of the same individual.[59]

Another way to response to Quine's problem about Ortcutt, the response that I am going to adopt, would be to deny with Quine (1956) that Ralph's beliefs really are (internally) inconsistent. The most straightforward way to go about this in the possible worlds semantics we are working in, is to follow Quine (1956)[60] and to judge a *de re* belief attribution like (12b) as being true iff Ralph has a description in mind (i) that actually fits Ortcutt, and (ii) whose instantiation is a spy in each of Ralph's belief-worlds. In this way, two *de re* belief attributions like (10) and (11) no longer lead to a contradiction, because

[58]For motivation, see Lewis (1986), and Stalnaker (1986).

[59]I concede, though, that I don't consider any of the three arguments as being conclusive.

[60]Or better, the way Kaplan (1969) formalizes some of Quine's remarks.

there might be two descriptions that actually fit Ortcutt, but do not denote the same individuals in all worlds compatible with what Ralph believes.

However, Kaplan (1969) noted that Quine's requirement for 'a conception of the individual the belief attribution is about' would make *de re* belief attributions too easily true. It is counterintuitively predicted that the *de re* belief attribution *Ralph believes of Ortcutt that he is the shortest spy* is true just because Ortcutt actually is the shortest spy and Ralph (*de dicto*) believes that the shortest spy is the shortest spy. Kaplan concluded that it is not enough for a *de re* attribution that the agent has a name or conception that happens to fit the individual whom the belief attribution is about in the actual world. Fit is not enough; the agent has to be *acquainted* with the individual whom the belief attribution is about because of some *causal relation*. In this way the problem of *the shortest spy* is resolved. Kaplan's solution is different from Russell's, because for Kaplan it is possible that a subject can be acquainted with an object in two different ways, such that he doesn't know that a single individual is the source of the two acquaintance relations. In this sense he follows Quine; however, since there is a sense in which Ralph does and a sense in which he does not believe of Ortcutt that he is a spy, we don't have to conclude that Ralph has inconsistent beliefs.

According to Kaplan, a belief can be about an individual, in our case Ortcutt, if two conditions are satisfied. First, the agent must have a representation of the individual; and second, this representation must be causally connected to the actual individual the belief is about — that is, Ortcutt. By the first condition, we can explain why the agent is disposed to perform certain actions that involve Ortcutt; by the second condition, we can explain how he came to have beliefs about Ortcutt. But we have seen above that these two conditions have to be satisfied not only in cases of *de re* belief ascriptions. Beliefs are always dependent on the environment, but in the case of *de re* belief, the causal relation, the *acquaintance relation*, is a very specific one. According to the causal-pragmatic account of intentionality, in all cases where a system represents or is about something else, the two conditions for aboutness demanded by Kaplan should be fulfilled. My conclusion is that Kaplan's analysis of *de re* belief attributions is not ad hoc, but part of a very general strategy to explain the notion of intentionality.

We have seen above that Kaplan's analysis of *de re* belief attributions fits the causal pragmatic explanation of intentionality. The content of what somebody believes should be explained in terms of counterfactual dependencies that hold, under certain normal conditions, between the belief state of the agent and his environment. This is the case both when we look at content from the agent's point of view, and when the relevant belief is really *about* the actual referent of a term used to characterize the agent's belief. Still, there is an important difference. In a case in which the belief is not really about the actual referent of the term, as with Bert's *arthritis*,[61] the agent's mental state tends to be sensitive not just to the actual referent of the term, but to everything that superficially looks like

[61] Cf. section 11.2.

it. In a case in which the agent's belief is really about the actual referent of the term, in our case Ortcutt, the beliefs of the agent according to which we can explain those actions of his that involve Ortcutt are sensitive primarily to facts about the real Ortcutt, and not to individuals that have a lot in common with Ortcutt.

In this section we have argued that to account for Quine's problem of *de re* belief attributions, we have to allow agents to be acquainted with the same individual in several ways. Technically this means that within possible worlds semantics we must assume that one actual individual might have *two* different *representatives* in the counterfactual worlds that help to represent Ralph's belief state. This can be accounted for by making use of a *counterpart theory*.

1.12.2 Externalism and Counterpart theory

Although it is obvious that this latter response is inconsistent with Kripke's view that proper names should be treated as rigid designators, this does not necessarily mean that it is also incompatible with the observations and arguments made by Kripke (1971, 1972). After all, the proposal to treat proper names rigidly need not be the only implementation consistent with his own observations and arguments. Kripke argued quite convincingly that it doesn't make much sense, if I say 'Suppose Nixon had lost the election', to ask whether a man in a counterfactual world resembles Nixon enough to be his counterpart in this counterfactual world. The argument in favor of heacceitism is quite convincing, but as I have suggested above this doesn't mean that the *actual* Nixon has to exist in all metaphysically accessible worlds, because heacceitism is compatible with a non-descriptive counterpart theory. The argument in favor of heacceitism also does not mean that *all* possible worlds are being considered when you make the *supposition* that Nixon has a property that he actually does not have.[62]

I think that it is a good idea not to make the metaphysical accessibility relation a universal accessibility relation, and I believe that this follows from the information-theoretic account of content, and is compatible with crucial observations made by Kripke (1972). As I argued above, when we check the truth of necessity statements in the actual world, we consider only counterfactual worlds in which the relevant normality conditions of the actual world hold. Twin Earth stories make it very clear that these normality conditions are *contingent*, and do not hold in all worlds. Twin Earth, in particular, will not be a world in which these normality conditions hold, and thus will not be considered to determine the truth value of a necessity statement in the actual world.[63] How are the normality conditions determined? Following Kripke, we can say that proper names like *Nixon*, common nouns like *cat, water*, and adjectives like *hot, yellow* etc. somehow have an actual extension, and that features of the members of this extension determine

[62] cf. Stalnaker (1981), and Muskens (1989).

[63] What if we want to evaluate a necessary statement in Twin Earth? In that case we only look at worlds where the normal conditions of Twin Earth hold.

the features that are essential to the individual, kind, or phenomenon.[64] For necessity statements we consider only counterfactual possibilities where there is an individual, a substance, or a phenomenon that has these essential features. It is irrelevant what these objects or phenomena are called in these counterfactual possibilities.[65]

No matter how we determine the normality conditions relevant to the checking of necessity statements, once we assume that the metaphysical accessibility relation is not the universal relation, we need no longer assume that cross-identification is *always* a matter of strict identity. Worlds that help to characterize a belief state need not be stipulated as counterfactual situations in which the referents of referential expressions are the same as in the actual world; it might be *counterparts* of the actual referent in this counterfactual world.

1.12.3 Counterpart theory

According to counterpart theory, the domains of different worlds are disjoint. To determine which beliefs a has about d in w, or which modal properties d has, we don't look at which properties d itself has in other worlds, but rather at the properties the counterpart(s) of d has in these other worlds.

Counterpart theory can be formalised in several ways. On Lewis' (1968) formalization, a formula like $\Box Rab$ would be true in a world if in every world containing counterparts of a and b, every counterpart of a bears the relation R to every counterpart of b. But, as noted by Hazen (1979), this gives rise to (at least) three closely related problems. First, it allows both $\Box Rab$ and $\neg\Box\exists x Rax$ to be true in the same world, because there might be metaphysically possible worlds in which the actual referent of b has no counterpart. Second, neither $\forall x, y[x = y \rightarrow (\Diamond x \neq y \leftrightarrow \Diamond y \neq y)]$ nor $\forall x, y[x = y \rightarrow \Box x = y]$ is predicted to be valid. Third, Hazen argues that the death of Caesar could, according to Lewis' theory, not be *essentially* of Caesar:

> Suppose in some possible worlds there were two counterparts of Caesar, living in opposite hemispheres of the globe. Each might be related appropriately - by dying it - to some counterpart of the death of Caesar, but neither could be related appropriately to the other's death. Thus neither counterpart of the death of Caesar is of all the counterparts of Caesar; so, if Lewis were right, the death of Caesar could not be essentially of Caesar. (Hazen, 1979, p. 329)

Hazen and Stalnaker (1986) have come to the conclusion that we should count a formula like $\Box Rab$ only true in a world, *not* if in every world containing counterparts of a

[64] See also Stalnaker (1979).

[65] Van Fraassen (1977) argues that the same holds for physical (logical, etc.) laws. A sentence is physically (logically, etc.) necessarily true if it is true in all physically (logically, etc.) accessible worlds. A counterfactual world is only physically (logically) accessible if all elements of the suitably chosen (?) set of so-called *law sentences* that hold in the actual world also hold in this counterfactual world. Thus, a sentence is physically (logically, etc.) necessarily true if it follows from the set of law sentences.

and b, every counterpart of a bears the relation R to every counterpart of b, *but* if for every relevant way of picking out counterparts, it holds that for every world in which a and b both have a counterpart, the counterpart of a bears the relation R to the counterpart of b. In other words, we should quantify not over counterparts, but rather over ways of picking out counterparts. Moreover, each way of picking out counterparts will really be a *function* from individuals and worlds to individuals (or representatives) in that world.

If c is such a counterpart function, c is a function that takes an individual, d, and a world, w, as arguments, and has an individual in the domain of w as its value. This latter individual might be called the *counterpart* of d in w with respect to c, $c_w(d)$. Formally this means that sentences are not only interpreted with respect to a context and index world, but also with respect to a counterpart function.[66] That is, a token of an atomic formula like $P(t)$ is interpret with respect to triples like $\langle w, w', c\rangle$ and is interpreted as follows:

- $[[\underline{P(t)}]]^{w,w',c} = 1 \quad \text{iff} \quad [[\underline{t}]]^{w,w',c} \in I_{w,w'}(\underline{P})$

Tokens of individual terms will be interpreted in terms of a counterpart function and the object denoted by $[\underline{t}]^{w,w'}$:

$$[[\underline{t}]]^{w,w',c} \quad = \quad c_{w'}([\underline{t}]^{w,w'})$$

To determine the object denoted by $[\underline{t}]^{w,w'}$, we first have to see what kinds of terms we have. For simplicity I have limited myself here to two kinds of terms, (i) tokens of demonstratives, and (ii) terms fronted by the diagonalisation operator †. For these two kinds of terms, $[\underline{t}]^{w,w'}$ is determined as follows:

$$\begin{array}{lll} [\underline{t}]^{w,w'} & = & \text{the utterer of } \underline{t} \text{ in } w, \text{ if there is one, and if } \underline{t} \text{ is a token of } I \\ & & \text{(and so on for the other demonstratives)} \\ & = & [\underline{t'}]^{w',w',g}, \text{ if } t = \dagger t' \text{ for some term } t' \end{array}$$

Belief attributions are interpreted with respect to counterpart functions as follows:

- $[[\underline{Bel(t, A)}]]^{w,w',c} = 1 \quad \text{iff} \quad \forall w'' \in K([[\underline{t}]]^{w,w',c}, w') : \ [[\underline{A}]]^{w,w'',c} = 1$

where $K(a, w)$ denotes the set of worlds compatible with what a believes in w.

Notice that according to the above interpretation rule, belief attributions are interpreted with respect to a single counterpart function that assigns each individual and world at most a single counterpart of the individual in this world. But if in some of Ralph's belief worlds the actual Ortcutt has *two* counterparts, which one do we refer to by a belief attribution like *Ralph believes that Ortcutt is a spy*? According to von Stechow (1984), and Stalnaker (1988), among others, which representation or counterpart we refer to depends not so much on the belief state of the agent itself, as on the intention of the speaker and

[66] Forgetting about variables for the moment.

on the conversational situation in which the belief attribution is made.[67] The counterpart/representative we refer to depends on the issue which one of Ralph's representations of Oscar is most salient in the relevant conversational situation. Formally this means that pragmatically speaking there will be a *unique* most salient counterpart function with respect to which the sentence should be interpreted.

Notice that when we assume that it depends on context which representation is relevant for the analysis of the *de re* belief attribution, there might be conversational situations where we might truly say that Ralph *doesn't* believe of Ortcutt that he is a pillar of society, which indeed seems to be in agreement with the facts.[68] If I only had given half of the story, and only told you that Ralph saw a man with a brown hat who behaves suspiciously, the belief attribution seems to be true.[69]

Although it seems clear that what is normally *communicated* by a *de re* belief attribution depends partly on the communicative situation, it is questionable whether it also determines the *truth value* of the belief attribution. This is suggested, in particular, by the following example as given by Richard (1983), which indicates that a *de re* belief attribution is already true if the agent believes the proposition expressed by the embedded sentence under *at least one* representation of the individual that the belief attribution is about:

> Consider A — a man stipulated to be intelligent, rational, a competent speaker of English, etc. — who both sees a woman, across the street, in a phone booth, and is speaking to a woman through the phone. He does not realize that the woman to whom he is speaking — B, to give her a name — is the woman he sees. He perceives her to be in some danger — a runaway steamroller, say, is bearing down upon her phone booth. A waves at the woman; he says nothing into the phone. [...] If A stopped and quizzed himself concerning what he believes, he might well sincerely utter:
>
> (3) I believe that she is in danger.
>
> but not
>
> (4) I believe that you are in danger.
>
> Many people, I think, suppose that [...] [these sentences] clearly diverge in truth value, (3) being true and (4) being false. [...] But [this] view [...] is, I believe, demonstrably false. In order to simplify the statement of the argument which shows that the truth of (4) follows from the truth of (3), allow me to assume that A is the unique man watching B. Then we may argue as follows: Suppose

[67] Indeed, it seems only reasonable to assume that this question should be answered in the same way as we answered the similar question for Kripke's cases of Peter and Pierre in section 1.12.3.

[68] But, as noted by Kaplan (1969), this is not predicted by Quine (1956).

[69] This approach towards *de re* belief attributions has been worked out in much detail in Aloni (2001). Staying close to my dissertation in this chapter, I have chosen not to discuss this work further here. The reader is encouraged to read it, though.

> that (3) is true, relative to A's context. Then B can truly say that the man watching her — A, of course — believes that she is in danger. Thus, if B were to utter
>
> (5) The man watching me believes that I am in danger
>
> (even through the telephone) she would speak truly. But if B's utterance of (5) through the telephone, heard by A, would be true, then A would speak truly, were he to utter, through the phone
>
> (6) The man watching you believes that you are in danger.
>
> Thus, (6) is true, taken relative to A's context.
> But of course,
>
> (7) I am the man watching you
>
> is true, relative to A's context. But (4) is deducible from (6) and (7). Hence, (4) is true, relative to A's context. (Richard, 1983, pp. 439-441)

To account for such examples, I will assume that although *pragmatically* there (should) exist(s) a unique most salient counterpart function with respect to which *de re* belief attributions are interpreted, *semantically* speaking we should *existentially* quantify over counterpart functions; A token of a sentence A is true in $\langle w, w' \rangle$ iff it is true with respect to at least one counterpart function (where C is the set of counterpart functions):[70]

- $[[\underline{A}]]^{w,w'} = 1 \quad \text{iff} \quad \exists c \in C : \ [[\underline{A}]]^{w,w',c} = 1$

In this way I propose to account for the intuition Richard (1983) pointed to that although what is communicated by a *de re* belief attribution is very context dependent, the belief attribution is still already true if the agent believes the proposition expressed by the embedded sentence under *at least one* representation of the individual the belief attribution is about.

In Appendix A I will give a more detailed formulation of our combination of two-dimensional modal logic and counterpart theory.

1.13 Info states, counterparts, and diagonalisation

In this chapter I have defended a causal information-theoretic analysis of content, and concluded that thus belief states should be represented in a rather coarse-grained way. I have argued that the problems this coarse-grained modelling posed for the analysis of belief attributions could be solved by adopting a three-way strategy; (i) taking notice at the *context dependence* of belief attributions, (ii) making use of *diagonalisation* in a two-dimensional theory of meaning, and (iii) using *counterpart theory*.

[70] This is somewhat different from what was proposed by Stalnaker (1986). Whereas I assume that we should *existentially* quantify over counterpart functions, Stalnaker proposes to use *supervaluation* to go from pragmatics to semantics.

For the analysis of *de re* belief attributions, I have argued that belief attributions in general should be interpreted with respect to counterpart functions. Let C be the set of all counterpart functions. We can say that the counterpart functions represent the way agents are *acquainted* with the objects they have beliefs about. It seems reasonable to assume that agents are only acquainted with a limited number of individuals, and that they can be acquainted with the same individuals in different ways. Now we can assume that there exists a set $C_a(w)$, a subset of C, which determines which individuals agent a has beliefs about in w: $BA_a(w) \stackrel{def}{=} \{d \in D(w) : \exists c \in C_a(w) \ \& \ \forall w' \in K_a(w) : c_{w'}(d) \neq *\}$, where $D'(w)$ is the domain of w', and $K_a(w)$ denotes the set of worlds compatible with what a believes in w. Once we have the sets $C_a(w)$ and $BA_a(w)$, we can also determine *how* the agent is acquainted with each individual of $BA_a(w)$; i.e. which *belief objects*, represented by individual concepts (functions from possible worlds to individuals), he associates with these individuals: $BO_a(w) \stackrel{def}{=} \{\lambda w'.c_{w'}(d)|K_a(w) : \ c \in C_a(w) \ \& \ d \in BA_a(w)\}$, where $f|K$ is the restriction of f to K.

Should we now say, then, that instead of representing the belief state of a in w by a set of possible worlds, we should rather represent it by a pair like $\langle K_a(w), C_a(w)\rangle$, where $K_a(w)$ is a set of worlds, and $C_a(w)$ the above discussed set of counterpart functions? I think we could, if we make an additional constraint on models. Normally we want to assume that agents know their own mind. The constraint that $\forall v, w \in W[v \in K_a(w) \rightarrow (K_a(v) = K_a(w))]$[71] is normally enough to encode the assumption that agents have introspective access to their beliefs; if they (do not) believe something, they believe that they (do not) believe it. However, once we represent belief states by pairs of the form $\langle K, C\rangle$, this is not enough anymore. The above constraint does not guarantee that if Ralph believes of Ortcutt that he is a spy, he also must believe that he believes of Ortcutt that he is a spy. Not only the set of doxastic accessible worlds should be the same in each belief-world, but also the set of belief objects. To account for this we can demand of the model that for every agent a, and world w, the following condition holds:

$$\forall u, v \in K_a(w) \rightarrow (K_a(u) = K_a(v) = K_a(w) \ \& \ BO_a(u) = BO_a(v) \supseteq BO_a(w))^{72}$$

If we represent belief states in the way suggested above, together with this constraint on models, this representation comes close to the representation of *anchored* beliefs in Kamp (1990). Then the question arises in what ways agents can form beliefs about individuals, or in terms of Kamp (1990), which relations give rise to anchored beliefs? Kamp suggests three such relations: Visual perception, memory, and the forming of a new belief in response to an utterance which contains a direct referential expression. I agree that in all those three ways agents can form beliefs about individuals, but I also think that it is much easier to form

[71] In terms of accessibility relations, this would mean that the doxastic accessibility relation for each agent would be *transitive* and *euclidean*.

[72] I assume that $BO_a(w)$ can be a subset of $BO_a(v)$, because agents can believe that they have beliefs of more individuals than they actually have.

beliefs about individuals by means of communication. We don't have to accept assertions in which a, what is traditionally called, direct referential expression occurs: normal *indefinites* and *pronouns* will do. This, I believe, is one reason why insights of modern theories of discourse, like for instance Kamp's (1981) Discourse Representation Theory, and Heim's (1982) File Change Semantics, are relevant for the more traditional issue of how to account for *de re* belief attributions. The reason is that under a particular interpretation of these theories the presupposition states used in them can be said to represent the information that participants of a conversation have of the individuals the conversation is about.[73] As a result, these presupposition states should be represented in basically the same way as I have suggested to represent belief states above.

I have suggested that belief states should be represented as being structured around belief objects. But doesn't that mean that, in the end, belief states should not be represented in terms of possible worlds only? No, it does not! The reason is that there is not only a relation between these modern discourse theories and *counterpart theory*, there is, or there should be, also a direct connection between these theories and *diagonalisation*. According to the causal/historical theory of reference, the referents of certain terms used by the speaker are determined by the 'causal' relations the speaker bears to the world. In this chapter I assumed that this is the case for proper names, but in the next chapter it will be argued that this also holds for most (other) uses of anaphoric expressions. And, just as agents might be unclear about what the referent of a proper name is because they are unclear about the origin of the relevant referential chain, agents might be unclear what the referent of a pronoun is because they are unclear about the causal origin of the relevant anaphoric chain. In this chapter I have argued that the first kind of unclarity for referential chains should be modelled by diagonalisation, and in the next chapter it will be shown that on a particular re-interpretation of the above mentioned modern theories of discourse these theories can be said to model the second kind of unclearness by diagonalisation too. What this suggests, is that for a lot of cases we can explain what the counterpart of an actual individual in another world is by means of diagonalisation, and thereby make a connection between the two techniques I crucially used in this chapter. Notice that by making this connection, i.e., by explaining at least *epistemic* counterparthood in such a causal way, we associate certain (descriptive) information with the counterpart of a certain individual in another world, and thus we would be able to represent belief states in terms of possible worlds only after all. The information associated with a belief object will typically not be an eternal description, however. Normally it will be information that involves a particular *token*, or *occurrence*, of an expression.[74]

[73] See the next chapter for this particular way of interpreting (a variant of) these theories.

[74] In the next chapters I will not stress my claim that we should be able to represent information states in terms of possible worlds only, but it should be clear that I always believe this is possible.

Chapter 2

Referential and Descriptive Pronouns

2.1 Introduction

Is it relevant to semantics whether the speaker has a certain individual 'in mind' by his use of the indefinite in a discourse like

(14) a. *A man* is walking in the park.

b. *He* is whistling.

and if so, how? On the one hand, according to Chastain (1975) and Donnellan (1978), among others, it is relevant both to the proposition expressed by the sentence in which the indefinite occurs, e.g. (1a), *and* to the propositions expressed by sentences with pronouns that take this indefinite as its syntactic antecedent, e.g. (14b). On the other hand, according to proponents of standard dynamic semantics like Kamp (1981), Heim (1982), and Groenendijk & Stokhof (1991), and to a neo-Russellian like Neale (1990), it is not (dynamic) semantically relevant at all: the object the speaker has in mind is at most important for pragmatics. In the first part of this chapter I will argue for a third option, originally proposed by Kripke (1977) and Lewis (1979b), and recently defended by Stalnaker (1998b), according to which speaker's reference is relevant to semantics, but only through pronominalisation. That is, it is truth-conditionally irrelevant for (14a), the proposition expressed by the sentence (or clause) in which the indefinite occurs, but is truth-conditionally relevant for (14b), the proposition expressed by a later sentence with a pronoun that takes an indefinite as its syntactic antecedent.

Recent theories of discourse representation are quite successful in accounting for anaphoric dependencies across sentential boundaries. But these theories face some problems, both conceptual and empirical. The information states used in these theories contain more than just truth-conditional content, because of their crucial use of *discourse referents*. The question arises of what this extra content could be, and what these discourse referents stand for. We would like to explain the status of discourse referents; they should not just be a tool for determining the truth-conditions of sentences that are interpreted with respect to

this information state (cf. Zimmermann, 1997). Given that pronouns are (interchangeable with) *definite* expressions, it seems reasonable to conjecture that a discourse referent is normally the hearer's representation in the informational or presuppositional state of the *speaker's referent*, as introduced by the speaker by his use of an indefinite description. On this view, pronouns are normally *referentially* used, referring back to the *unique* and specific object the speaker has had 'in mind' by his use of the antecedent indefinite. Unfortunately, this is not the way pronouns and discourse referents *are* and *can* be thought of according to the above theories of discourse representation, which all treat pronouns essentially as bound variables *existentially* closed at the text level. I will argue, however, that there is *empirical* evidence for a referential analysis of most occurrences of personal pronouns in sentences like (14b), and that by means of *diagonalisation* (Stalnaker, 1978) and the use of *hypothetical* reference-contexts in a *two-dimensional* theory of reference, such an analysis can be pushed much further than many have supposed. In fact, this analysis gives almost, although not quite, the same predictions as the above-mentioned theories.

Of course, not all indefinites are specifically used, and we can still sometimes refer back to these indefinites with singular pronouns. However, for these uses of pronouns a notion of *uniqueness*, or *exhaustivity* also seems to be involved. Only on this assumption can the *definiteness* of all singular pronouns be explained. So, I will be proposing that the singular pronouns that take indefinites that are not specifically used as their syntactic antecedents, and which are represented by existential quantifiers, should be treated as *descriptive* pronouns, referring (if at all) to the *unique* individual, or the *exhaustive* set of individuals, that satisfies the description recoverable from the sentence in which the antecedent indefinite occurs.

The remainder of this chapter can be roughly divided into two parts. The first part (until section 6) is about the referential use of pronouns, the second part about the descriptive use. The first part is organised as follows. In section 2, I discuss some classical approaches to anaphora. In section 3, I argue on the basis of some empirical phenomena for a referential analysis of anaphoric pronouns, and for taking seriously the notion of speaker's reference in dynamic semantics. I also show how the view that speaker's reference influences truth conditions can be formalised in terms of an occurrence analysis. In section 4 I explain how this analysis can be related to the standard dynamic systems. In section 5, I discuss the problem of how to explain successful communication when anaphoric pronouns are thought of as referential expressions. I argue that this can be done when we think of dynamic semantics, through *diagonalisation*, as an extension of the traditional two-dimensional theory of reference. I then show how we can explain the *status of discourse referents* in information states by means of this occurrence analysis, which provides a more satisfying explanation than standard dynamic systems can. In section 6, I discuss the relation between my analysis of referential pronouns and Donnellan's famous analysis of referentially-used definite descriptions.

In the second part of the chapter, I argue that a singular pronoun can sometimes be appropriately used in the 'main' context even when it does not refer to the speaker's referent of the indefinite. This will motivate postulating the existence of descriptive pronouns, for which I will provide a systematic implementation in a dynamic semantics. Next, I show how *functional pronouns* – needed to handle, for instance, Karttunen's (1969) notorious paycheque examples – can be accounted for in our dynamic framework in such a way that the definiteness constraint on singular pronouns can still be satisfied.

In the final section, I will suggest that by making use of descriptive and functional pronouns we can also account for the universal effect of donkey sentences in a descriptive way, thereby making a strong connection between the specific/unspecific use of indefinites on the one hand, and the referential/descriptive use of pronouns on the other.

To this chapter belong two appendices where some formalities are discussed; one concerning the formalization of standard dynamic semantics, and the other the analysis of referential pronouns as proposed in section 3 of the chapter.

2.2 Some classical approaches to anaphora

According to scholastic approaches to indefinites and anaphora, pronouns can refer back to indefinites because indefinites are referential expressions. The indefinite refers to that object that the speaker intends to refer to by the use of the indefinite. Moreover, if a speaker uses a referential expression in his utterance, the proposition expressed by this utterance is object-dependent. Geach (1962) has criticised this account. If John intends to refer to d by his use of the indefinite *an S*, and wants to say of d that he is P, even though d is not, John is not saying something false when he claims *An S is P*, according to Geach, if there actually is an S that is P. In order not to make such a prediction, according to Geach, it is better to represent an assertion like *An S is P* semantically simply by an existential formula, $\exists x[Sx \wedge Px]$. The specific/unspecific distinction belongs to pragmatics, which should be kept separate from semantics. To handle pronouns, we should follow Quine's insight and treat them as bound variables. A sequence of the form *Some S is P. It is Q* should, according to him, be translated as $\exists x[Sx \wedge Px \wedge Qx]$.

But there are well-known problems with this latter assumption. First, it leads to the unnatural consequence that we can interpret a sentence with an indefinite or other anaphoric initiator only at the end of the whole discourse: incrementality is given up. Second, if we want to interpret the pronouns in a donkey sentence like *If a farmer owns a donkey, he beats it* as bound variables, it seems we have to represent the indefinites in the antecedent as universal quantifiers to get the truth conditions right. But then it seems we have to give up compositionality. We cannot treat indefinites in all contexts in the same way. Finally, sometimes we cannot even get the truth conditions right by assuming that all pronouns should be treated as bound variables. This was shown by Gareth Evans (1977). Evans convincingly argued that we sometimes denote by our use of a pronoun all

the relevant objects by which the antecedent sentence is verified. Thus, in a sequence of the form *Some S are P. They are Q*, the pronoun *they* goes proxy for the description *(all) the S such that P*.[1] Such pronouns he called *E-type pronouns*; I will sometimes also call them *descriptive pronouns*. Thus, I will call a pronoun an E-type pronoun if it goes proxy for the description recoverable from its antecedent clause.

The existence of E-type pronouns was argued for on the basis of the following kind of example:

(15) Tom owned some sheep and Harry vaccinated *them*.

According to a Geachian analysis of this sentence, we learn that Harry vaccinated some sheep that Tom owned if we accept what is expressed by the sentence; what we seem to learn, though, is that Harry vaccinated *all* of the sheep that Tom owned. The latter reading is predicted if the pronoun *them* is analysed as an E-type pronoun.

I consider it undeniable that E-type pronouns do exist; but that doesn't mean that all pronouns are E-type pronouns. There is one obvious reason for this. The pronouns occurring in sentences like

(16) Every man loves *his* cat, and

(17) Each woman liked the man who gave *her* a rose.

seem to function like the bound variables of quantification theory. Indeed, since Evans (1977), proponents of the E-type approach normally make a distinction between *bound* and *unbound* pronouns, claiming that such a distinction can be made on purely syntactic grounds; and propose that only unbound pronouns should be treated as E-type pronouns. A pronoun P is a bound pronoun, and treated as a bound variable, roughly if it is anaphoric on and thus bound by a quantifier Q, only if P is located inside the smallest clause containing Q (Neale, 1990, p. 171).[2]

However, if we use the term *unbound pronoun* in the above sense, it seems that not even all unbound pronouns go proxy for the definite or universal noun phrase recoverable from the antecedent clause and should be treated as E-type pronouns. Consider the following example due to Dekker (1994):[3]

(18) Yesterday, John met some girls. *They* invited him to their place.

[1] Evans (1977) claimed that the pronoun *rigidly refers to* (all) the S such that P. See Neale (1990) for a motivation of the interpretation I have chosen. I will give some additional motivation later. Still, I agree with Evans's claim that (many) unbound pronouns are referring expressions. Nevertheless, I will argue that these pronouns are not E-type pronouns.

[2] For a more specific syntactic characterisation, see Evans (1977) and Neale (1990). In my later discussion of epistemic *might*, I argue that there is indeed something to the distinction between bound and unbound pronouns as characterised by these authors.

[3] For similar examples, see Sommers (1982) and Kamp & Reyle (1993).

In this case, we don't want to say that *they* needs to stand for all girls John met yesterday. If we want to say that the pronoun is going proxy for a description recoverable from its antecedent, the relevant description should not be definite or universal, but *indefinite*.[4] The description would be *Some girls that John met yesterday.* To treat the pronoun as an abbreviation of an indefinite description also seems to be needed to get the right reading of a sentence like

(19) Socrates owned a dog, and it bit Socrates.

It seems that (19) can be true if there was a dog that Socrates owned and it bit him, although at the same time there was also another dog that he owned that did not bite him. But claiming that the pronoun is an abbreviation of an *indefinite* description would be very implausible. Pronouns are *definite* expressions:[5]

> 'It' [is] a definite singular term whether its antecedent is or not. 'He', 'she', and 'it' are definite singular terms on a par with 'that lion' and 'the lion' [...] The three compound sentences 'I saw a lion and you saw that lion', 'I saw a lion and you saw the lion', and 'I saw a lion and you saw it' are interchangeable. Such use of a definite singular term dependently upon an indefinite antecedent [...] makes no distinction between a pronoun such as 'it' and a singular description such as 'the lion'. (Quine, 1960, p. 113)

Should we therefore treat all unbound pronouns as abbreviations for *definite* descriptions recoverable from their antecedent clauses after all? There might be a way to get rid of the unwelcome resulting uniqueness prediction that arises in some cases,[6] although the prospects look rather dim. First, it doesn't seem to be a very natural strategy to explain away 'apparent' counterexamples to the uniqueness assumption by assuming that the domain of quantification is always selected in such a way that the uniqueness effect is reached after all. Second, sometimes even domain restriction doesn't help. This is shown by donkeys in bishop's clothing:

(20) If a bishop meets another bishop, he blesses him. (Heim, 1990)[7]

If pronouns are treated as recoverable *definite* descriptions, it seems to be impossible to select the domain in the correct way. As argued above, giving up the assumption that pronouns are definite expressions doesn't seem to be natural.

But if a singular pronoun cannot be treated as a definite description that (in extensional contexts) refers to (all) of the object(s) that verify the antecedent sentence, how then can a pronoun be treated as a definite expression?

[4] See van der Does (1994).

[5] See also Kadmon (1990).

[6] For early discussion, see Evans (1977); see Neale (1990) and Heim (1990) for some more recent ones.

[7] Attributed to Kamp and to van Eijck . See section 2.11 for further discussion.

The answer given by Kamp (1981), Heim (1982), and more recent proponents of dynamic semantics like Groenendijk & Stokhof (1991), Chierchia (1992), and Dekker (1993) is familiar by now: treat anaphoric pronouns simply as bound variables, interpret indefinites dynamically in such a way that they introduce new objects that are available for reference, and assure that in the case of negation universal quantification over assignment functions or sequences of individuals is involved. Anaphoric pronouns can as such be treated as definite noun phrases, because the possibilities with respect to which the pronouns are interpreted are finer-grained entities than possible worlds; namely, world-assignment pairs. From now on I will denote all dynamic theories simply by CCT, for *Context Change Theory*. (In Appendix B, I will formulate standard CCT as it is given in Dekker (1993).)

2.3 A referential analysis of anaphoric pronouns

I have argued above that, intuitively, anaphoric expressions should be thought of as *definite* expressions, but cannot in general be treated as abbreviations of definite, or universal, descriptions recoverable from an antecedent clause. It seems that the definiteness of pronouns is accounted for in CCT, because in each possibility in which a singular pronoun is interpreted, the pronoun will 'refer' to, or denote, at most one particular individual. But this way of looking at CCT is rather misleading. In fact, it is more appropriate to say that in all versions of CCT, including Discourse Representation Theory (DRT; Kamp, 1981), File Change Semantics (FCS; Heim (1982), Dynamic Predicate Logic (DPL; Groenendijk & Stokhof 1991), and most explicitly in van der Does (1994), the pronouns in sentences (14b), (18) and (19) are treated as abbreviations of *indefinite* descriptions.[8]

However, this treatment of pronouns cannot account for the *definiteness* of pronouns after all. Moreover, as the following data suggest, a pronoun does not simply go proxy for the indefinite description recoverable from the antecedent clause. In the cases below, the pronouns should receive a more specific interpretation than the theories mentioned above can offer.

It is commonly assumed that the phenomenon of *pronominal contradiction* shows that anaphoric pronouns can at least sometimes be used referentially. When John asserts (21a), Mary can react by saying (21b):

(21) a. John: *A man* is running through the park.

b. Mary: *He*'s not a man, but just a boy.
And *he* is not running, but just walking.

In these cases it is clear that the pronoun cannot be used as an abbreviation for the indefinite description *a man who is running through the park*. It is more reasonable to

[8]To be more precise, in the discourse *An S is P*. He *is Q*. He *is R*, the first occurrence of *He* goes proxy for the indefinite description *An S who is P*, while the second occurrence goes proxy for *An S who is P and Q*.

assume that the pronoun is used referentially, referring to the *speaker's referent* of John's use of the indefinite.

The following example,[9] which shows what I will call the *specificity problem*, suggests that pronouns are, in fact, generally used in this way. If John says (22a), it would be odd for him to reply to Mary's question (22b) by saying (22c)

(22) a. John: *A man* called me up yesterday.

b. Mary: Did he have a gravel voice?

c. John: That depends: if *he* called in the morning *he* did,
but if *he* called in the afternoon, *he* did not.

if two men called John up yesterday and he knows this. It not easy to see how this phenomenon can be explained if it is assumed that pronouns should simply be treated as variables bound by dynamic existential quantifiers. As noted by Dekker (1997), it also seems clear that the phenomenon cannot be explained in terms of the classical entailment relation between what the speaker believes and what he says; (22a) and (22c) are wrongly predicted to be fine given that John knows that two men called him up yesterday, one in the morning and one in the evening. To explain that (22c) cannot be used appropriately in its most straightforward reading in such a context, a *more specific* relation than entailment is needed to account for the intuition that John just wants to talk about one of the two men. A natural explanation can be given if it is assumed that for the use of a pronoun the speaker must have a specific object 'in mind'.[10] Such a more specific entailment relation can be given when we make *possibilities finer-grained* than in ordinary CCT; in that case a distinction can be made between the two situations where John had two different individuals 'in mind' for the use of the indefinite.

On the assumption that pronouns are normally used referentially, we can also explain the frequently observed distinction between the discourse (23a) and the single sentence (23b):

(23) a. There is a doctor in London. *He* is Welsh.

b. There is a doctor in London *who* is Welsh.

which, according to the standard account, are predicted to be equivalent. The distinction is this one: for the use of the personal pronoun in the discourse (23a), the speaker must have a specific individual 'in mind' that the second sentence with the personal pronoun is about; whereas no individual need be 'in mind' to ensure the acceptability of the single sentence (23b), in which a *relative* pronoun is used.[11]

[9] This example came up in a discussion with Paul Dekker and Ede Zimmermann.

[10] See also Dekker (1997).

[11] A further argument for my claim that pronouns not c-commanded by their antecedents normally refer back to *specifically* used indefinites comes from *intentional identity* attributions. This will be discussed in the next chapter.

According to the causal/historical theory of reference, the referents of certain terms used by the speaker are determined by his intentions; and the content of the intentions depend, in turn, on 'causal' relations that the speaker has with the world. Normally this is assumed for proper names only; in this chapter, however, I will argue that this is also true for most (other) uses of anaphoric expressions.[12] The main claim of the first part of this chapter is that we can account for the range of phenomena dynamic semantics can account for if we assume that most anaphoric pronouns in the original fragment of DRT/FCS/DPL are used referentially. This is because a pronoun normally picks up the relevant speaker's referent of its antecedent indefinite,[13] the object the speaker has 'in mind', which is understood as the object that was 'causally responsible' for his use of this token, or occurrence, of the expression.[14]

In an influential discussion of the proposal to treat pronouns as referential expressions, Heim (1982, §1.3) argues, partly on the basis of the asymmetry in acceptability between (24a) and (24b), against a referential treatment of pronouns:

(24) a. John owns *a donkey*. Mary beats *it*

b. John is a donkey-owner. *Mary beats *it*,

She argues that this asymmetry cannot be predicted on the basis of the truth conditions of the first sentence in each discourse and the surrounding circumstances alone, because what seems crucial is how each sentence is worded. It seems that the difference can not be accounted for by means of the existence and absence of speaker's reference in the first and second discourses respectively, either; even if the speaker had a specific donkey in mind in the second discourse, the use of the personal pronoun would still be odd. Heim observes that if pronouns are treated as variables bound by 'text-scope' existential quantifiers associated with explicitly mentioned indefinites, the asymmetry can be explained; and in the later chapters of Heim (1982), she argues that this latter approach is in fact the way to go.

The same argument applies, according to Heim (1982) and Kamp (1988), to the contrast in acceptability between the following examples:[15]

(25) a. Exactly *one* of the ten balls is not in the bag. *It* is under the sofa.

[12]This analysis has an antecedent in Sommers (1982), where it is argued that proper names should be thought of as 'special duty' pronouns: pronouns that can be used in more than one conversation.

[13]For plural pronouns the speaker's referent is not a unique individual, but rather the exhaustive set of individuals that the speaker has in mind for his use of the antecedent indefinite.

[14]Note that because of the way I understand 'speaker's reference', my analysis does not give rise to a problem that Heim (1982) observes for Kripke's (1977) analyses. Heim (p. 17) argues that a pronoun can take the indefinite *a dog* in *A dog has been rummaging in the garbage can* as its 'syntactic antecedent', although the speaker has no idea which dog is responsible for the mess the speaker sees. On my analysis, the indefinite has a speaker's referent: the individual responsible for the mess the speaker sees, which indirectly caused the speaker's use of the indefinite.

[15]The example is attributed to Partee.

b. Exactly *nine* of the ten balls are in the back. *?*It* is under the sofa.

The second discourse here is also odd, according to Heim and Kamp, because no explicit indefinite has been used. So, all that counts for the explicit use of a pronoun is whether an indefinite has been explicitly used in the previous discourse. Discussing examples (22a) - (22c), we have already noted that this cannot be a *sufficient* condition. Now I want to argue that it is also not a *necessary* condition. The explicit use of an indefinite is also not a necessary condition for the appropriate use of an anaphoric pronoun because, as has been observed by many authors, the pronoun *it* can be used appropriately in (25b) when the speaker makes it clear that he is interested in the tenth ball (by looking for it for a moment), or that he has the tenth ball in mind.[16]

What this suggests is that it is not so much the explicit use of an indefinite that counts, but rather that the speaker has made it clear to the hearer(s) that he has a specific individual 'in mind'.

But why, then, is it at least normally the case that the speaker can use a personal pronoun to 'refer' back to an explicitly used indefinite? Why are (24b) and (25b) normally so much worse than (24a) and (25a)? The reason, I wish to suggest, is that it is a *speech convention* among language users, and thus known to be speech convention, that when a speaker uses an indefinite explicitly, he *normally* has a specific individual in mind. Because, as stressed by Stalnaker (1998b), one kind of information hearers can receive from the use of a sentence is that the sentence was uttered and which particular words were used, the hearer will assume that the speaker has a specific individual in mind when he uses an indefinite. As a result, thinking of pronouns as referential expressions can explain the above asymmetry. The reason is that in the first sentences of (24a) and (25a) but not of (24b) and (25b), an indefinite is explicitly used, and thus only in the former case can a pronoun be appropriately used.[17]

To account for this referential use of pronouns, we have to assure that the interpretation of a pronoun is the speaker's referent of its antecedent indefinite. Moreover, we have

[16]See, for instance, the following scenario sketched by Neale (1990, p. 209):

> Suppose I have ten pet mice, one of whom is called 'Hector'. Hector is always getting out of the cage in which I keep all ten mice, and whenever he does so he goes and hides under the sofa. I open up the cage and begin counting mice: "One, two, three,..." When I reach 'nine' I turn to you and with a knowing look I say,
>
> (i) I put all ten mice in the cage an hour ago, and there are only nine here now.
>
> Knowing Hector's habits, you might then reply,
>
> (ii) I bet *he*'s under the sofa again.

[17]I do not want to suggest that pronouns can never refer back to incorporated nouns, cf. van Geenhoven (1996), but in distinction with van Geenhoven I do believe that this can only happen with *descriptively* used pronouns, which 'refer back' to all objects satisfying the relevant descriptive material. Normally you cannot refer back to such incorporated nouns by *singular* pronouns, because normally it can not be presupposed that there is only a *unique* object that satisfies this description.

to guarantee that the possibility with respect to which the specifically used indefinite is interpreted assigns a *unique* individual (if there is one) to the representation of this indefinite; it depends exclusively on the possibility which unique individual is introduced by each occurrence of an indefinite. As a result, we have to assume that possibilities should contain more information than possibilities contain in standard dynamic semantics; they should also indicate what the speaker's referent is, if there is one, for each occurrence of a (specifically used) indefinite.

In standard dynamic semantics, a distinction is made between information about the *subject matter* of conversation, real-world information, and information about the values of discourse referents. I have argued in this section that a possibility should contain at least one more piece of information: information concerning the identity of the speaker's referents of specifically used indefinites. Intuitively, facts determining what the speaker's referent of an occurrence of an indefinite is are *facts* about the *world*. It is *tokens* of indefinites that have speaker's referents; and different tokens of the same type of indefinite might have different speaker's referents. For formal reasons, however, I will not go into all of the complexities of a token analysis, and will differentiate between facts about the world relevant to the *subject matter* of conversation and facts determining the speaker's referents of occurrences of indefinites. Instead of using a token analysis according to which it is *tokens* of indefinites that have speaker's referents, determined by facts about the world, I will make use of an *occurrence* analysis, assuming that it is *occurrences* of indefinites that have speaker's referents, and let an additional function determine the identity of the speaker's referent of an occurrence of an indefinite. I will assume that occurrences of specifically-used indefinites are represented by *indexed eta terms*, of the form '$\eta r_n A$'; and that a possibility is not a world-assignment pair, but rather a world/*reference context*/assignment triple, where a reference context is a function from indices to individuals. Because each specifically-used indefinite will be represented by an indexed eta term, it will be clear for each world/reference context/assignment triple which individual, if any, is referred to by such an indefinite. Although formally it will be the case that an occurrence of a specifically used indefinite can have different speaker's referents in a world, because a world can form a possibility with many reference-contexts, we can assume that for each world there exists a *distinguished* reference-context such that this will never be the case for such possibilities.

An occurrence of a sentence of the form *A man is walking in the park* is represented by something like $WiP(\eta r_n Man)$. In order to assure that in following sentences the speaker's referent of the indefinite becomes available for reference for anaphoric pronouns represented by the discourse marker r, we have to enrich the partial assignment function. If $\langle w, c, g\rangle$ is the possibility with respect to which the occurrence of the above sentence is interpreted, the enriched assignment function also assigns a value to r, namely the speaker's referent of the indefinite *a man* in possibility $\langle w, c, g\rangle$, $c(n)$. If the pronoun *he* in the following sentence *He is whistling* takes the indefinite *a man* as antecedent, we can say that this

following sentence is true in $\langle w, c, g[^r/_{c(n)}]\rangle$ if and only if $c(n)$ is whistling in w. If we limit ourselves to singularly-used indefinites and pronouns, we have three kinds of terms: (indexed) eta terms, discourse referents, and ordinary variables. These terms represent indefinites, anaphoric pronouns, and relative pronouns, respectively; and are interpreted as follows:

- $[[t]]^{w,c,g} = g(t)$, if t is a variable or discourse referent,
  $= c(n)$, if $t = \eta r_n P$ and $c(n) \in I_{w,c,g}(P)$

The analysis to this point sounds very much like what has been proposed by Chastain (1975), Donnellan (1978), and Fodor & Sag (1982). However, these earlier analyses claim not only that pronouns can refer back to speaker's referents of indefinites, but also that when an indefinite is used specifically, a sentence like *An S is P* is false if the speaker's referent of the indefinite is not an object with property P, although there is another object that is an S and is P. Although I wish to claim that specifically used indefinites come with a speaker's referent, I want to follow Kripke (1977), Lewis (1979b), and Stalnaker (1998b) in taking the speaker's referent of the indefinite to be semantically irrelevant to the interpretation of the clause in which the indefinite itself occurs.[18] That is, for the truth of a sentence of the form *An S is P*, only the *existential* information counts. But this doesn't mean that it is semantically irrelevant which object it is the speaker has in mind for his use of the indefinite, as is assumed by standard dynamic semantics. Just as Kripke (1977) argued, speaker's reference is relevant to semantics, but only through *pronominalisation*. But now, of course, the following question arises: How could we account for this *referential* analysis of pronouns, on the one hand, and the *existential* interpretation of indefinites, on the other? I will answer this question by giving a definition of the *truth of a discourse* defined in terms of an update function *Upd*, and a notion of *rigid truth*, which are defined *separately* from (although in the end they are mutually dependent on) each other.

Let us begin with the definition of $Upd(E, \langle w, c, g\rangle)$, which tells us how the partial assignment function g is enriched after the interpretation of expression E.[19] In this section I will discuss only the most important clauses, assume that all predicates are simple, and leave the formulation of the complete theory argued for in this section to Appendix C.

- $Upd(\eta r_n P, \langle w, c, g\rangle) = g[^r/_{c(n)}]$

[18] For the same reason I also don't want to use an E-type analysis that assumes that the interpretation of the indefinite quantifier is contextually restricted, such that the singular pronoun satisfies the uniqueness condition after all. Although my analysis, just like this modified E-type analysis, assumes that the uniqueness condition for singular pronouns is normally satisfied because of some additional contextual information, I believe that this additional information is semantically irrelevant to the clause in which the indefinite occurs.

[19] The idea to separate context change from determining truth conditions in a dynamic framework is not new. It can also be found in van der Does (1994) and Peregrin & Von Heusinger (1997). These authors do not argue for a referential analysis of anaphoric pronouns, though.

- $Upd(t, \langle w, c, g\rangle) = g$, if t is a variable or discourse referent
- $Upd(R(t_1, .., t_n), \langle w, c, g\rangle) = Upd(t_n\langle w, c, Upd(t_{n-1}, .., Upd(t_1, \langle w, c, g\rangle)..)\rangle)$
- $Upd(A \wedge B, \langle w, c, g\rangle) = Upd(B, \langle w, c, Upd(A, \langle w, c, g\rangle)\rangle)$

Now we can determine the notion of *rigid truth*, where we determine the truth of the sentence with respect to a possibility when we do not existentially quantify over the reference-contexts.

- $[[R(t_1, ..., t_n)]]^{w,c,g} = 1$ iff $\langle [[t_1]]^{w,c,g}, ..., [[t_n]]^{w,c,g}\rangle \in I_{w,c,h}(R)$ where $h = Upd(t_n\langle w, c, Upd(t_{n-1}, .., Upd(t_1, \langle w, c, g\rangle)..)\rangle)$
- $[[A \wedge B]]^{w,c,g} = 1$ iff $[[A]]^{w,c,g} = 1$ and $[[B]]^{w,c,h} = 1$, where $h = Upd(A, \langle w, c, g\rangle)$

Notice that although the speaker's referent of an indefinite like *a man* need not be a man, as required to account for pronominal contradiction examples, it follows by the interpretation rule of indefinites and atomic sentences that for a sentence like *A man is sick* to be (rigidly) true, it needs to be the case that there is a man who is sick.

Finally, we define the notion of *truth of sentence A in* $\langle w, c, g\rangle$, $\langle w, c, g\rangle \models A$, by existentially quantifying over the set, C, of reference contexts.

- $\langle w, c, g\rangle \models A$ iff $\exists c' \in C$ such that $[[A]]^{w,c',g} = 1$.

Notice that when A is interpreted after the sequence of sentences S_1 to S_n, A will be true with respect to initial context $\langle w, c, g\rangle$ iff there is a $c' \in C$ such that $[[A]]^{w,c',h} = 1$, where $h = Upd(S_1 \wedge ... \wedge S_n, \langle w, c, g\rangle)$. Thus, our definition assures that an indefinite is always interpreted existentially, and a referentially-used anaphoric pronoun always refers back to the speaker's referent of its antecedent indefinite.

It is easy to see that both sentences of the discourse *A man is walking in the park. He is whistling*, as represented by $WiP(\eta r(Man))$. $Whistling(r)$, are now predicted to be true in $\langle w, c, g\rangle$ iff there exists a man who is walking in the park, $I_w(Man) \cap I_w(WiP) \neq \emptyset$, and the speaker's referent of the indefinite is whistling, $[[\eta r(Man)]]^{w,c,g} \in I_w(Whistling)$, just as we wanted. Notice that because the speaker's referent of the indefinite of the first sentence need not actually be *walking* in the park, a second speaker might react by saying that *he* is not walking, but running, which shows that we can also account for pronominal contradiction examples.

Of course, the referential analysis of pronouns raises a serious problem: namely, the problem of *donkey sentences*. Although we have suggested that the referential analysis of pronouns can be pushed further than is usually assumed, the question remains how it can account for donkey sentences like (20), repeated here as (26):

(26) If *a bishop* meets *another bishop*, *he* blesses *him*.

This, of course, is one of the examples for which dynamic semantics was originally invented. The problem is that in the above sentence the indefinites seem to have no speaker's referents; and even if it is assumed that the definiteness of some pronouns should be explained by the assumption that they are used descriptively, these donkey sentences still cannot be accounted for in any straightforward way. This is because these pronouns cannot be treated as abbreviations for the *definite* description *the unique bishop that meets another bishop* recoverable from the antecedent clause, for the obvious reason that there is no such bishop.

Fortunately, we can account for donkey sentences in a manner very similar to that in which standard dynamic semantics accounts for them: we can say that in donkey sentences, too, the indefinites are used specifically and the pronouns referentially, but the speaker's referents of the indefinites are determined not with respect to the *actual* reference context, but with respect to all *hypothetical* reference-contexts of set C.[20] That is, we can account for the universal effect of donkey sentences by assuming that the analysis of negation and adverbs of quantification (and ordinary quantifiers) involves a quantification over those *hypothetical* reference contexts that make the same facts true with respect to the subject matter of the conversation as in the world, or possibility, being considered, but not with respect to the facts that determine the speaker's referents of indefinites. The way to do this is to assume that negation and (adverbial) quantifiers are treated as *intensional* operators, in that they allow part of the context, i.e. the reference context, to shift (where ADV and DET stand for any adverb or quantifier, and $[ADV]$ and $[DET]$ for their usual interpretation):[21]

- $[[\neg A]]^{w,c,g} = 1$ iff $\neg\exists c' \in C:\ [[A]]^{w,c',g} = 1$
- $[[ADV(A,B)]]^{w,c,g} = 1$ iff $[\ ADV](\{Upd(A,\langle w,c',g\rangle):\ c' \in C\ \&\ [[A]]^{w,c',g} = 1\},$ $\{Upd(A,\langle w,c',g\rangle):\ c' \in C\ \&\ [[A \wedge B]]^{w,c',g} = 1\})$
- $[[Det_x(A,B)]]^{w,c,g} = 1$ iff $[Det\ ](\{d \in D:\ \exists c' \in C\ \&\ [[A]]^{w,c',g[^x/_d]} = 1\},$ $\{d \in D : \exists c' \in C\ \&\ [[A \wedge B]]^{w,c',g[^x/_d]} = 1\})$[22]

On such an 'intensional' treatment of negation and adverbial quantifiers as shifters of reference contexts, a conditional donkey sentence – represented either as in DPL in terms of conjunction and negation by $\neg(Own(\eta r_n \hat{x} Fx, \eta s_m \hat{y} Dy) \wedge \neg Beat(r,s))$, or as in DRT and FCS by fronting the two arguments by an implicit adverb of quantification,

[20] But see the last section of this chapter for a rather different account of donkey sentences that does not involve hypothetical reference contexts.

[21] Note that Kaplan (1989) would call such an 'intensional' treatment of negation and (adverbial) quantifiers *monstrous*.

[22] Note that in this way I assume that sentences with quantified noun phrases always receive a *weak*, or *selective*, reading; while sentences with adverbs of quantification always receive a *strong*, or *unselective*, reading. As I will explain below, however, I have no principled reason to make this distinction. I just want to show how both analyses could be implemented.

$Always(Own(\eta r_n \hat{x} Fx, \eta s_m \hat{y} Dy), Beat(r, s))$ – gets the usual *universal*, and *unselective* reading.[23] The reason is that we look at *all* reference-contexts that assign to occurrences of indefinites specific individuals.[24] Because we keep the world fixed, to determine the truth of the consequent of the conditional in world w, we have to look at all farmer-donkey pairs that stand in the 'own' relation in w.

Moreover, if we say that $Upd(\neg A, \langle w, c, g \rangle) = g$ (and something similar for (adverbial) quantified phrases), we can guarantee that negations and (adverbial) quantifiers figure as plugs with respect to anaphoric binding, just like in classical DRT/FCS.

Until now I have been implicitly assuming that all indefinites should be represented by eta terms. However, I believe that some occurrences of indefinites are used *quantificationally*, and so cannot be referred back to with referentially-used pronouns. Quantificationally-used singular indefinites will be represented with the one-place quantifier '$\exists$'. If P is a one-place predicate, I will say that $\exists P$ is a sentence. The one-place predicates need not be simple, but can also be *complex*. That is, if A is a sentence, I will say that $\hat{x}A$ is a complex one-place predicate. Complex predicates and existential sentences are interpreted as follows:

- $I_{w,c,g}(\hat{x}A) = \{d \in D : [[A]]^{w,c,g[x/d]} = 1\}$
- $[[\exists P]]^{w,c,g} = 1$ iff $I_{w,c,g}(P) \neq \emptyset$[25]

Notice that once we have complex predicates, these complex predicates might themselves contain eta terms that introduce discourse referents into the discourse. Indeed, in our formal analysis we should be able to account for this. For ease of exposition, however, I will neglect these formal problems here, and leave the details to Appendix C.

2.4 Comparison with standard dynamic semantics

The theory sketched above is close to the dynamic semantic theories developed by Kamp (1981), Heim (1982), and Groenendijk & Stokhof (1991). All of these analyse donkey sentences in a similar way, and take indefinites to introduce into the discourse objects to which we can refer back in later sentences. But there are at least three important differences between these approaches and the one just sketched. The first is a result of the *truth definition* of a sentence that I have given. Whereas indefinites and *pronouns* are basically treated as bound variables existentially closed at the text level in standard dynamic semantics, a pronoun refers back to the *unique* speaker's referent of an *occurrence* of an indefinite according to the truth definition of sentences given above. As a result, the

[23] For the analysis of *asymmetric* readings of conditional donkey sentences, see Appendix C.

[24] But we have to guarantee, of course, that there are enough reference contexts in the model, one for each farmer-donkey pair.

[25] Of course, when the indefinite *some man* in *Some man is walking through the park* is only existentially used, the sentence has to be represented by $\exists \hat{y}[Man(y) \wedge WiP(y)]$.

sentences in the discourse (23a), *There is a doctor in London. He is Welsh* is represented by $E(\eta r_n \hat{x} DLx).\, Wr$ (where 'E' is the existence predicate defined by $\hat{x}\exists\hat{y}[x = y]$), need not have the same truth conditions according to my analysis as the sentence (23b), *There is a doctor in London who is Welsh*, represented by either $E(\eta r_n \hat{x}(DLx \wedge Wx))$ or $\exists\hat{x}[DLx \wedge Wx]$, as is predicted to be the case on a standard dynamic semantics account. We don't predict the two to be equivalent when the speaker's referent of the relevant occurrence of the indefinite *a doctor* is not Welsh, although another doctor in London is. Note also that the truth conditions of the latter two representations of (23b) differ if it is 'rigid truth' that counts, but are identical with respect to the non-rigid notion of truth. Thus, whether an indefinite is specifically used or not is relevant only for pronominalisation, just like I argued above.

The second difference is that the approach sketched here can account for the phenomena of *pronominal contradiction* and the *specificity problem* discussed above, because it is assumed that the speaker must have a specific object 'in mind' for his use of indefinites and pronouns. When John asserts (22a), *A man called me up yesterday*, we predict that John makes the specific individual that is the speaker's reference of the indefinite *a man* available for reference for pronouns and (other) short descriptions. This speaker's reference is the specific individual he has in mind for his use of the indefinite, and because he refers back to this speaker's reference with a pronoun it would normally be odd for him to answer the question (22b), *Did he have a gravel voice?* by saying (22c), *That depends; if he called me up in the morning he did, and if he called me up in the afternoon, he did not*, because it can be assumed that the speaker had in mind either the one who called him up in the morning or the one who called him up in the afternoon. According to my analysis, if the speaker says (21a), *A man is running through the park*, the speaker's reference of the indefinite need not be walking in the park, or even be a man. This is in accordance with the facts, because if the hearer knows which individual the speaker has in mind, he can respond by saying (21b), *He is not a man, but just a boy. And he is not running, but just walking.* In this way we can account for the phenomenon of pronominal contradiction.[26]

Of course, some proponents of standard dynamic semantics might be skeptical of my use of the notion of speaker's reference within a *semantic* theory. They might think that the notion of speaker's reference is irrelevant to semantics, even for determining the *truth conditions* of later sentences in which anaphoric pronouns occur. On this view, the notion might at best be relevant to pragmatics. I have two responses to such skeptics: (i) Even if the notion of speaker's reference is relevant only to pragmatics, the resulting pragmatic analysis would be, I claim, very close to my *semantic* analysis; (ii) The phenomenon of *pronominal contradiction* shows that the notion of speaker's reference is at least sometimes relevant for determining the truth conditions of a later sentence in which an anaphoric pronoun occurs.

[26]Dekker (1997) has independently proposed a very similar analysis to account for pronominal contradiction and the specificity problem. But see the next footnote.

But suppose that the notion of speaker's referent were *not* relevant to truth-conditional semantics. Suppose, moreover, that we did not have to deal with the phenomenon of pronominal contradiction. What, then, could an integrated semantic/pragmatic theory of anaphoric relations look like? For the skeptics who make these assumptions I have a very simple proposal: assume that the analysis I have given above is just a pragmatic account, and think of a semantic analysis of anaphoric relations as an *abstraction* of this pragmatic analysis.

In the theory that I have sketched above, possibilities contain more information than the possibilities used in standard dynamic theories; they also contain the information about what the speaker's referents are of particular *occurrences* of indefinites. This extra information is responsible for the main differences between my proposal and the standard accounts that I pointed out above. But, of course, it is possible to abstract away from this extra information; and if we do so, what results is (truth-conditionally equivalent to) the standard dynamic theory. That is, when we don't want to say that pronouns should refer back to speaker's referents, we can define the truth of sentence A after sequence S as follows: A is true* after $S_1...S_n$ in $\langle w, c, g\rangle$, $\langle w, c, g\rangle \models^*_S A$, iff there is a $c' \in C$ such that $[[S_1 \wedge ... \wedge S_n \wedge A]]^{w,c',g} = 1$. As a result, the so-called (conjunctive) donkey-equivalence between *There is a doctor in London. <u>He</u> is Welsh* and *There is a doctor in London <u>who</u> is Welsh* holds again, just as in standard dynamic systems. Thus, I am not claiming that proponents of standard dynamic systems are saying anything wrong, but only that they aren't saying enough: they should take the notion of speaker's reference more seriously than they actually do.[27]

The occurrence analysis given above allows for a finer-grained treatment of discourses and, and as I have argued, dynamic semantics is in need of a finer-grained analysis of this kind to account for some *empirical* phenomena that are problematic for standard accounts. However, the most important difference between standard dynamic semantics and the alternative that I have given above is I believe, not empirical, but rather *conceptual*: it is about the status of discourse referents.

[27] Although the account of pronominal contradiction and the specificity problem discussed above is close to Dekker's (1997) proposal, he does not think of standard dynamic semantics as an abstraction of the kind of analysis I have given here. Whereas I assume that my 'pragmatic' analysis is basic and that the standard dynamic semantic picture can be derived by making possibilities less fine-grained, Dekker assumes that standard dynamic semantics is basic and that the notion of 'speaker's reference' can be captured by building pragmatics *on top of* the standard account. Of course, this difference need not be substantial, as long as Dekker assumes that capturing the notion of 'speaker's reference' requires him to make the possibilities finer-grained than in ordinary CCT. Unfortunately, this does not seem to be the way he wants to account for speaker's reference; and as a result, I do not see how speaker's reference is really accounted for in Dekker's analysis – especially when we assume that presupposition states should be *introspective*, as I will argue later.

2.5 Discourse referents and diagonalisation

2.5.1 Unclear reference and successful communication

In the previous sections, I argued that pronouns normally refer back to the speaker's referents of their antecedent indefinites, and showed how this view can be implemented. But the resulting treatment seems to have an unwelcome consequence which is avoided on standard accounts. I have argued above that pronouns are referential expressions, referring back to the speaker's referent of their antecedent indefinites. A common assumption in the philosophy of language is that in determining the referents of referential expressions, one can represent a context by an n-tuple of objects, and that it is clear to both speaker and hearer what this context is. This latter assumption is based on a Gricean conversational maxim: speakers ought to assume that hearers have enough information to determine what proposition they have expressed. If the hearer fails to recognize what object is referred to by a referentially-used expression, then he cannot determine what proposition is expressed by the speaker, who thus violates the conversational maxim. It seems to follow that if some anaphorically-used pronoun is treated as a referential expression, the speaker has to presuppose that the hearer can recognize what object the speaker is intending to refer to by the pronoun. Otherwise the hearer will not understand what is meant by the sentence in which the pronoun occurs. Unfortunately, this is commonly not the case, and the hearer cannot tell which object the speaker has been intending to refer to with the indefinite or pronoun. But how, then, can communication be successful?

It seems that in order to account for successful communication we have to give up the assumption that pronouns are all, in these cases, being used referentially and refer back to the speaker's referent of its antecedent indefinite. Below I will argue, however, (i) that we are not, in fact, forced to accept this conclusion, since we *can* account for successful communication on the assumption that many such pronouns are being used referentially; and (ii) that we *should* explain these pronouns in this way in order to explain the status of discourse referents in information states used to analyse discourses.

2.5.2 Bridging the gap by diagonalisation

According to the causal/historical theory of reference, the referents of certain terms used by the speaker are determined by the 'causal' relations that the speaker bears to the world. Normally this is assumed only for proper names and demonstratives; in this chapter, however, I have argued that this also holds for most (other) uses of anaphoric expressions. But just as agents might be unclear about what the referent of a proper name or demonstrative is, because they are unclear about the origin of the relevant referential chain, agents might also be unclear about the referent of a pronoun, because they are unclear about the causal origin of the relevant anaphoric chain. In the previous chapter, I followed Stalnaker (1978) in claiming that we can describe how successful communication is achieved despite the un-

certain reference of proper names and demonstratives by means of diagonalisation. In this chapter, I want to argue that the reference for pronouns and successful communication should be bridged by diagonalisation too.

Ideally, a referential expression is used only when it is clear to the hearer what the expression refers to. It is clear, though, that ideal conditions do not always obtain. If the speaker says something and the hearer disagrees, there might be *two* reasons for this disagreement. First, the hearer might have understood what the speaker has said, but he disagrees with the speaker on the facts the discourse is about. Second, speaker and hearer might agree about these latter facts, but disagree because the hearer thinks that the speaker has said something different from what he has actually intended to say. The latter situation might obtain if the speaker uses a referential expression. These two different reasons for disagreement can be accounted for in the *two-dimensional* theory of reference proposed by Kaplan (1989) and Stalnaker (1970b).[28] The reason is that in this theory a conceptual distinction is made between two kinds of facts: (i) facts about the *subject matter* of conversation, and (ii) facts about the *conversational situation* itself. What is expressed by a sentence, then, might depend on the facts of the conversational situation.

Suppose Hans says *I will see you at 10 o'clock tomorrow* in a conversation with Ede and Paul. Although Hans intends to refer to Ede by his use of the demonstrative pronoun *you*, Paul might react by saying *No, because I will take the train to Amsterdam this evening*. In this case, Paul need not disagree with Hans about the facts relevant to the conversation's subject matter, but he just misunderstood what Hans has intended to say because Hans's use of *you* has been accompanied by an unclear pointing gesture. If we say that reference contexts represent facts about the conversational situation, we can think of a reference context in this simple situation as a possible referent of the demonstrative pronoun *you*. Clearly, there are two possible referents, Ede and Paul. In a two-dimensional theory of reference, we can represent what Hans has said as a function from reference contexts to the proposition expressed by *I will see you tomorrow* in this reference context: $\{\{w \in W |\ a \text{ will see Hans tomorrow in } w\} : a = \text{Ede or } a = \text{Paul}\}$. Of course, this function from reference-contexts to propositions is formally a Kaplanian (1989) *character*.

Whereas in a Kaplanian two-dimensional framework sentential *types* denote *characters*, functions from reference-contexts to propositions, in the Stalnakerian counterpart it is sentential *tokens* that express *propositional concepts*, functions from worlds to propositions. The idea is that worlds play two roles: one role, an *index role*, for which only facts about the subject matter of conversation count; and a *context role*, for which only facts about the conversational situation count, and of which facts about linguistic and speech conventions are important ingredients. The idea is that the same token of an expression (or a counterpart of it) might have a different interpretation in a different (context)world, because the facts about the conversational situation might be different in this other world. For instance, although Hans actually referred to Ede by his use of the demonstrative *you*,

[28] See chapter 1.

Paul thought that Hans has referred to him, because he has been mistaken about certain *facts* of the conversational situation.

If w_0 is the actual world and $\underline{A}$ a token of sentence A, the actual *horizontal* proposition expressed by this token of A – that is, the set of *index worlds* where what is expressed by the token of A with respect to the actual conversational situation is true – is determined as follows: $\{w' \in W|\ [[\underline{A}]]^{w_0,w'} = 1\}$. Although it is normally the horizontal proposition that the speaker intends to express, the hearer, as we have seen, doesn't always know which one this is because he does not know the relevant facts about the conversational situation. But even if the hearer doesn't know which horizontal proposition is expressed by the sentence, the information that he receives from the sentence can still be modelled as a proposition, a set of possible worlds, if we make use of a token analysis.

A context should contain not only the information available for the interpretation of context-dependent utterances, but also the information accepted by speaker and hearer about the *subject matter* of the conversation. Because in two-dimensional modal logic both kinds of facts are treated as facts about the world, we can represent a context set, S, by a set of worlds. Any element of S might, as far as the hearer can tell, be the actual world that makes true everything that is presupposed, both about the subject matter of the conversation and about the conversational situation itself. If any element of S might, as far as the hearer can tell, be the actual world, he might update this information state S after accepting the utterance by eliminating any world w in S in which what is expressed in w is false in w. This new information state is $\{w \in S|\ [[\underline{A}]]^{w,w} = 1\}$, and is what Stalnaker (1978) has called the *diagonal proposition* expressed by $\underline{A}$ with respect to S.

Stalnaker (1978) proposes that each time we can assume that the speaker assumes that it is unclear for the hearer which horizontal proposition is expressed by a token of a sentence, we should reinterpret what is said and assume that it is the diagonal proposition that the speaker has intended to communicate. According to this diagonalisation solution, successful (enough) communication does not require there to be a unique individual that is the referent of a referential expression in all worlds consistent with what is presupposed. Just as we can explain by means of diagonalisation how the identity statement *Hesperus is Phosphorus* can successfully and informatively be used in a communicative discourse, we can also explain the successful use of an anaphoric pronoun. Successful communication does not require that there be, in the former case, a unique individual that is the referent of the names *Hesperus* and *Phosphorus* in all worlds consistent with what is presupposed; or that there be, in the latter case, a unique individual that is the referent of the pronoun in all worlds consistent with what is presupposed. Thus, by means of diagonalisation we can explain how successful communication can be achieved despite the uncertain reference of anaphoric pronouns.

It is important to realise that in order to extend the diagonalisation strategy from proper names and demonstratives to anaphoric pronouns, we *have* to assume that anaphoric pronouns are referential expressions. The reason is that diagonalisation requires that in

each world consistent with what is presupposed, the relevant expression must have a *unique* referent.[29] This condition is not met in theories like DRT, FCS or DPL. However, it is if it is assumed that most (singular) pronouns are used referentially, and refer to the *unique* individual that is the speaker's referent of the antecedent indefinite in the world under consideration.

In this section, I have assumed that worlds contain information both about the subject matter of conversation and about the conversational situation itself. As a result, it is possible to represent a context and a diagonal by a set of possible worlds. In my formal analysis, however, I have assumed that worlds contain information just about the subject matter of conversation and not about the conversational situation itself. It follows – forgetting about the introduction of individuals for the moment – that a context has to be represented by a set of world/reference context pairs, and that the *diagonal* expressed by A with respect to context set S is $\{\langle w, c\rangle \in S : \ [[A]]^{w,c} = 1\}$. It is important to keep in mind that the two are not crucially different. A possible world, both in a two-dimensional analysis and in the intuitive sense of the word, is the same as a world/reference context pair in my formal analysis.

2.5.3 The status of possibilities and discourse referents

If we assume that most anaphoric pronouns are used referentially, the result is that it becomes unclear, generally speaking, which horizontal proposition is expressed by a sentence with an anaphoric pronoun. Thus, for context change due to sentences in which anaphoric pronouns occur, diagonalisation becomes the rule rather than the exception. I will argue, in fact, that context change in dynamic semantics should normally be thought of in terms of diagonalisation.[30] Although this is not the way dynamic semantics is normally understood, I believe that it is the way we *should* understand it if we want to give a *non-representational* account of discourse interpretation. One of the reasons for this is that, in this way, possibilities used to represent presupposition states need not be finer-grained than possible worlds in the intuitive sense of the word.

According to the *functional* analysis of attitudes, an agent stands in a certain attitude relation to a proposition, if by means of this relation, together with the assumption that the agent is rational, we can explain the agent's behaviour. Attitudes are dispositional, or functional, states of a rational agent; and these states are individuated by the role that they play in determining the behaviour of the agent who is in such a state. This picture (see Stalnaker, 1970b, 1973, 1974) suggests that presupposition should also be thought of as a propositional attitude: we have to know what the speaker is presupposing in order to

[29] Thus, there is no requirement that $\exists x : \forall w : x$ is the speaker's referent in w of the relevant expression, but only that $\forall w : \exists x : x$ is the speaker's referent in w of the relevant expression.

[30] Note that, as a result, context world and index world are never different, and so we don't need possibilities with more than one world.

explain his behaviour when he is engaged in a conversation.[31] The alternative possibilities that help to represent what the speaker is presupposing are the relevant alternatives consistent with what the speaker assumes is commonly assumed, and with respect to which we have to judge the informativity and acceptability of the speech acts made by speakers.

Although in theories like DRT and DPL *truth* is an important concept, in the dynamic semantic theories of Heim (1982), Dekker (1993), and Groenendijk et al. (1996), the notion of *information* is most important. Sentences are not interpreted with respect to a fixed world and assignment function, but rather with respect to information states represented by sets of world-assignment pairs. An information state represents what is *presupposed* about the conversation, and contains information about both the subject matter of conversation (the 'world' information) and the values of discourse referents. Of course, the two kinds of information are connected: the values assigned to discourse referent r in world w introduced by the indefinite *a man* in *A man is walking in the park* must all be men walking in the park in w. What is important is that with respect to different world-assignment pairs in the context, pronouns will 'refer' to different individuals, because the discourse referent introduced by an indefinite might receive different values, even if we fix a world.

The possibilities in information states of the kind used in dynamic or update semantics should be finer-grained than possible worlds, so it is argued, because they represent something about the discourse going on in the actual world. But what exactly do these possibilities represent about this discourse? The fact that a certain noun phrase in the discourse has been used, you might think. But how can the information that a certain noun phrase has been used be enough to explain why pronouns can take these kinds of noun phrases as their syntactic antecedents? Because this is just a fact about how our language works, you might respond. The phenomenon that shows the *definiteness* of anaphora, as discussed in section 3 of this chapter, suggests that something more is needed: the existence of a discourse referent in an information state should represent not only the fact that an indefinite is used in the discourse, but also that the speaker has had a specific individual 'in mind' by his use of the indefinite. We have already seen that pronouns can normally be used appropriately only if the speaker has a specific individual in mind that he intends to refer to. Thus, it seems reasonable to assume that (the information associated with) a discourse referent in an information state of the kind used in dynamic semantics is the representation in this state of the fact that the speaker has had a particular individual 'in mind' for his use of the indefinite that has introduced this discourse referent into the discourse.

Something like this has been proposed by Zimmermann (1997) and Dekker (1997).[32] They both argue that discourse referents in the information states used in dynamic seman-

[31] With Stalnaker I will assume that presupposition as a propositional attitude is a more basic notion than the semantic presupposition relation between sentences or propositions triggered by specific lexical items; and that the latter relation should be explained in terms of the former. See chapter 4 for more on presuppositions.

[32] And in van Rooy (1997).

tics represent something about the speaker's actual intentions. I agree that these discourse referents should represent something about these intentions. In particular, they should normally be representations of the speaker's referents of the indefinites used in the discourse that the information state is a representation of. This suggests that presupposition states, representing information about specific speaker's referents, and belief states, representing information about objects that the belief is about, should be represented in a similar way (cf. Dekker (1997), van Rooy (1997), and Zimmermann (1997)).

It is generally agreed that for an agent to have beliefs about an individual, two conditions must hold: the *external* condition that this particular individual must have (partly) *caused* the agent's belief state in order for the agent to be in this particular state, and the *internal* condition that the agent believes that his representation is a representation of a particular individual. It is only reasonable to assume that for a presupposition state to have a representation of an individual, the same two conditions must hold. The external condition can easily be met by enriching the information state with an anchor, linking, or counterpart function between actual individuals and representations of these actual individuals (see Kamp (1990) and the previous chapter). To meet the internal condition, however, each possibility, I will argue, should assign to a discourse referent introduced by an indefinite the *unique* speaker's referent of this indefinite in this possibility.

In the earlier sections of this chapter, I assumed that speaker's intentions were relevant to semantics: what a pronoun refers to depends on the speaker's referent of the antecedent indefinite. Now I want to argue that if we want to take seriously the assumption that a presupposition state should be represented as a propositional attitude, we have to represent the information states of dynamic semantics differently from the way that they are normally represented: we should not allow contexts containing two or more possibilities with different assignment functions but the same possible world or, in our formal analysis, with the same world/reference context pair. The argument will be that a discourse referent in an information state should not only represent something about the speaker's intentions, but should also be *generally assumed*, or *presupposed*, to represent something about the speaker's intentions.

A belief state is usually represented by a set of possible worlds; and each of those worlds might, as far as the agent believes, be the actual world. To explain the actions of rational agents, it is normally assumed that believers know their own minds, i.e. have *introspective* access to their own minds; if an agent believes or doesn't believe something, he also believes that he does or doesn't believe it. I have argued above that presuppositions, just like beliefs, should be thought of as propositional attitudes, needed to explain the communicative actions of agents. But if speech is action, and if the appropriateness of the speech acts of agents is to be explained partly in terms of what they presuppose, we have to assume that the attitude of presupposition is also liable to introspection, so that if an agent presupposes something (about the discourse), he also presupposes that he presupposes this something, and if he doesn't presuppose something (about the discourse), he presupposes

that he doesn't presuppose it.

So, just as each world of a belief state might, as far as the agent believes, be the actual world, each element of the context (the possibilities consistent with what is presupposed) might, as far as the participants in a conversation assume for the sake of the conversation, be the actual possibility where the discourse is taking place. Remember that possible world semantics assumes that what somebody believes is a *fact* about the world. If presupposition is a propositional attitude, it is only natural to assume that what somebody presupposes in a possibility is also a *fact* about this possibility.[33] If a speaker makes an assertion that is accepted by the participants in a conversation, it is not only a *fact* about the actual possibility that the assertion has been made and that the truth of what is expressed by this assertion is presupposed. It is also a *fact* about the possibilities consistent with what is presupposed that the assertion has been made and that the truth of what is expressed by this assertion is presupposed (see Stalnaker, 1978, 1998b). Now suppose that in possibility α of a presupposition state, discourse referent r is assigned to object d and that S is the set of possibilities compatible with what is presupposed (by the speaker) in this possibility. It can be assumed that every possibility in S also assigns an individual to r, but this need not be d. Although in each possibility there should be a unique individual that is made salient by the indefinite that has introduced r, there need not be a unique individual presupposed to be this individual in order for the communication to be successful. Now the following question arises, whether or not α represents the actual possibility: What is represented about possibility α by the information associated with r in S?[34]

Suppose that possibility α is consistent with what is presupposed, and represents what a possibility represents in standard dynamic semantics. Suppose, in addition, that in the world of the possibility there are two men walking in the park, and that we are looking at an information state resulting from the update of an earlier information state with the assertion *A* man_r *is walking in the park*. The question that arises now is what the information associated with r in S (the context representing what is presupposed about the conversation in α) represents about $g(r)$, if g is the assignment of α. Proponents of standard dynamic semantics would claim that this question makes no sense; but, as Stalnaker (1998b) points out, and as the problems discussed earlier suggest, the question does, as least as regards the actual world/possibility, seem to make sense. It seems natural to answer this question for the actual world/possibility by looking at the specific individual the speaker has in mind in his use of the antecedent indefinite. That is, we would say as regards the actual world/possibility that the individual that an indefinite makes available

[33]That the possibilities used in dynamic semantics should contain the information that is presupposed in the conversation is something that I have learned from Fernando (1997) and Stalnaker (1998b) (see also Zeevat, 1997). But Fernando, at least, does not draw from this the conclusion that I will argue we should draw: that the pronouns analysed in dynamic semantics should be treated as referential expressions.

[34]We might say that while Zimmermann (1997) and Dekker (1997) ask this question only with respect to the *actual* world/possibility, we should also should ask this question with respect to every world/possibility consistent with what is presupposed.

for reference is the *speaker's reference* of the particular use of this indefinite. And if this is so for the actual world/possibility, why not, then, for worlds/possibilities compatible with what is presupposed to be the actual world/possibility, such as our α?

On this account a personal pronoun refers to the presupposed speaker's referent of its antecedent; there need not, however, be a unique individual that is presupposed to be the speaker's referent of its antecedent indefinite. As I argued earlier, successful communication requires only that in each world compatible with what is presupposed, there is a (unique most salient) individual available to which this pronoun or description may refer. The speaker presupposes such an individual by virtue of the fact that he has acted in a particular way – making it clear that he has introduced a speaker's referent into the discourse. In normal cases, the relevant act by which the speaker's referent is introduced is the speaker's use of an indefinite.

Even if the speaker knows which individual is responsible for his use of the indefinite, the hearer need not and may not know anyhing about this individual that the speaker has intended to refer to with his use of the indefinite. In general, it seems that a discourse referent is associated only with the information that it verifies the sentence in which the indefinite occurs and that the speaker has intended to refer to it by his use of a certain indefinite. Thus, the information that the hearer has about a speaker's referent and that the speaker can presuppose about it can be thought of as the *diagonalised speaker's referent*. This, I suggest, is the information associated with a discourse referent.

If we assume that $\langle w_0, c_0, d_1^0, ., ., ., d_m^0, C^0\rangle$ is the actual possibility, where $\langle d_1^0, ., ., ., d_m^0\rangle$ are the actual speaker's referents of the occurrences of indefinites that introduce discourse markers r_1 to r_m into the discourse and C^0 is what is presupposed in this actual possibility, we can represent C^0 as follows:

$$C^0 \quad = \quad \left\{ \begin{array}{cccccccc} \langle\ w_1 & c_1 & d_1^1 & . & . & . & d_m^1 & C^1\ \rangle \\ \langle\ w_2 & c_2 & d_1^2 & . & . & . & d_m^2 & C^2\ \rangle \\ & & & . & . & . & & \\ & & & . & . & . & & \\ & & & . & . & . & & \\ \langle\ w_n & c_n & d_1^n & . & . & . & d_m^n & C^n\ \rangle \end{array} \right\}$$

Here, each row represents what might be the actual possibility as far as is presupposed. My proposal that the presupposition state, C^0, is introspective means that it will be the case for each $i: \ 1 \leq i \leq n: \ C^i = C^0$.[35] In addition, each column in the above matrix associated with discourse referent r_j represents not only d_j^0 in C^0, but also d_j^i in each C^i, where $1 \leq i \leq n$. This means that the column represents not only the actual speaker's referent (if there is one) of the indefinite that has introduced r_j into the discourse, but also the *presupposition* that this represents the actual speaker's referent.

[35]This seems problematic for standard set theory, but I will argue in chapter 4 that this is not really the case if we model what is presupposed in terms of an accessibility relation.

If we assume that terms can introduce objects into the discourse and that what somebody presupposes is a fact about the world, we cannot analyse context change in an *eliminative* way by means of diagonalisation, as was suggested above.[36] A more complicated analysis is needed: we have to assume that when an assertion is accepted the world/possibility changes too; and that the *actual* world/possibility need not be an element of the set representing what is presupposed in this possibility. What we really need to do is to change the definition of '$Upd(A, \langle w, c, g, C\rangle)$' so that a possibility is updated with the terms introduced by A; and that it is *presupposed* that A is true and accepted and that the referents of the terms of A are also introduced. I will sketch such an analysis only in chapter 4 of this book. But the analysis will have the result that after the interpretation of, for instance, a sentence of the form $R(t_1, ..., t_n)$, the possibility is enriched by the objects introduced by the terms, and $R(t_1, ..., t_n)$ is presupposed after the update to be true *and presupposed*.

Although this way of representing presupposition states and the way that these states change is close to the notion of context and the update function used in standard dynamic semantics, there are some differences.

One difference concerns the effect of using an indefinite in standard dynamic semantics, the use of an indefinite normally adds uncertainty, because each world-assignment pair can be extended in several ways, given the uncertain reference of the indefinite. According to our analysis, however, the reference of an indefinite in each possibility is already determined. In this sense the standard theory is admittedly more natural: from the *hearer's* point of view, more possibilities are relevantly different after the update with the indefinite than before it. However, because on my analysis possibilities always contain *more* information than on the standard account, there is an easy way to solve this problem. If S denotes a set of world/reference-context/assignment triples, I can bring my notion of context into line with that of the standard account. This 'standardisation' of S, $s(S)$, is, of course, $\{\langle w, g\rangle : \exists c \in C : \langle w, c, g\rangle \in S\}$. The uncertainty that an indefinite in sentence A adds to the context can now be accounted for by comparing the standardisation of S with that of the context as results from the interpretation of A with respect to S. The intuition, then, that more possibilities are relevant after than before the use of an indefinite is accounted for if we distinguish possibilities after update whose differences were truth-conditionally irrelevant before update.

There are, however, two more significant differences between my account and the standard one. One is that the *status of discourse referents* in standard dynamic semantics is not clear, while in the theory I have just sketched it is. The other is that the standard

[36] And because presupposition is thought of as a propositional attitude, the following (where sentence A is interpreted with respect to context, or presupposition state, S) will not do either:

$$[[A]](S) \quad = \quad \{\langle w, c, Upd(A, \langle w, c, g\rangle)\rangle : \langle w, c, g\rangle \in S \quad \& \quad [[A]]^{w,c,g} = 1\}$$

.

account cannot explain the *definiteness* of pronouns, while I can. The reason for the first difference is that on my account each occurrence of a specifically used indefinite in the actual world/possibility has a unique speaker's referent. Although this world/possibility need not be consistent with what is presupposed, and although the speaker need not know which object is the speaker's referent of a specifically-used indefinite, it is this object that the discourse referent represents, and the world/possibility in question is the one about which things are presupposed. Thus, once we assume that the actual world/possibility, and (actual) speaker's reference count, we can say that a discourse referent represents the presupposed information about a certain speaker's referent. Moreover, by using diagonalisation and by assuming that presupposition is a propositional attitude, we can also say that a discourse referent is presupposed to represent something about the speaker's referent. To be a bit more concrete: let $\langle w, c, g, C\rangle$ be a possibility such that C is the context set that represents what is presupposed in this possibility. Now we can ask what information this context set C has about the value of a variable and what it represents about the possibility $\langle w, c, g\rangle$. To answer the first question, we can use the notion of *subject* as defined by Dekker (1993). Normally, assignment functions are seen as functions from discourse referents to individuals, but we might say that discourse referents are functions from possibilities to individuals. The information that context set C associates with discourse referent r, i.e. the subject of C associated with r, can now be defined as follows:

$$[r]_C \stackrel{def}{=} \begin{array}{l} \text{the function } f \in [C \rightarrow (D \cup \{*\})] \text{ such that} \\ \forall\langle w, c, h, C'\rangle \in C : f(\langle w, c, h, C'\rangle) \;=\; h(r) \end{array}$$

The question of what such a subject represents about possibility $\langle w, c, g, C\rangle$ can now be answered straightforwardly: the subject of C associated with r, $[r]_C$, is the representation in C of $g(r)$. If $g(r)$ is the speaker's referent of a specifically used indefinite in possibility $\langle w, c, g, C\rangle$, the speaker can presuppose that the information associated with this individual is $[r]_C$. Thus, we can say that discourse referents really stand for something, and are not just used to be able to interpret later sentences. As a result, the theory is truly *non-representational* in the sense of Zimmermann (1997). The reason that I can explain the *definiteness* of pronouns, and the standard account cannot, is that both in the actual world, and in (other) worlds consistent with what is presupposed, pronouns refer back to the *unique* speaker's referent of their antecedent indefinites.

2.6 Referential descriptions and propositional concepts

Note that given the way that I have represented possibilities in the last section, we can determine both the *diagonal* and *horizontal* propositions expressed by a sentence. We might say that (an occurrence of) sentence A expresses in possibility $\langle w, c, g\rangle$ the following

diagonal, $[\dagger A]$, and horizontal, $[A]$, propositions, respectively:[37]

$$[\dagger A](\langle w, c, g\rangle) \quad = \quad \{v \in W|\ \exists c', h : \langle v, c', h, C'\rangle \in C\ \&\ v, c', h \models A\}$$

$$[A](\langle w, c, g\rangle) \quad = \quad \{v \in W|\ v, c, g \models A\}^{38}$$

So, just as in any two-dimensional framework, we can determine the whole propositional concept associated with a sentence, and therefore the two kinds of propositions that might be expressed by that sentence. But then the question arises whether we ever need more than just the diagonal: do we ever need the whole propositional concept?

It seems that we do, because the speaker normally intends to express the horizontal proposition determined by the sentence. But proponents of the standard dynamic account will not be so easily convinced. Look at a case where the speaker uses a demonstrative pronoun. In an assertive utterance of *You are sick* the speaker certainly intends to express the horizontal and *object-dependent* proposition expressed by the sentence, a proposition which says something *about* a particular object. Still, looking only at the diagonal does not really seem problematic. But isn't it true that the diagonal proposition expressed by the above sentence is context- and thus object-independent? Yes, in principle this diagonal proposition is object-independent, but what is relevant is always the diagonal proposition expressed with respect to a special context. If all reference context/index pairs were relevant, the above claim wouldn't say much more than *the person to whom the speaker is speaking is sick*. In most conversational contexts, however, it is pretty clear to the participants in the conversation who the speaker and addressee are: the same individuals in all reference contexts of the context. It follows by the above claim that the same object-dependent proposition would be determined in all relevant reference contexts, and thus that the diagonal of the relevant propositional concept would be object-dependent, too.

Still, I agree with Stalnaker (1970b) that we need the whole propositional concept. Reference context/index pairs should not be merged into primitive points of reference such that propositions are considered to be functions from reference points to truth values.[39] The horizontal propositions expressed are of some independent interest, and to bring that out there has to be a functional difference between reference context and index. One reason that we sometimes need the full propositional concept is that we want to be able to make a distinction between the *a priori* and the *necessary*. If we always look only at the diagonal, in a sense we always look to see whether what is said is *a priori* true: we look at things

[37]Below I make the somewhat simplified assumption that only reference context c and assignment g are relevant to the determination of what is horizontally expressed by a sentence.

[38]Note that for both the diagonal and the horizontal propositions expressed by sentences, it holds that $Q(\eta r_n(P))$ and $\exists\hat{x}[Px \wedge Qx]$ determine the same proposition. Thus, whether an indefinite is specifically or unspecifically used is irrelevant to the proposition expressed by the sentence in which the indefinite itself occurs.

[39]See also Lewis (1980).

only from an *epistemic* point of view. But sometimes we want to know whether what is said is, for instance, *necessarily* true or not physically speaking, as in *I didn't have to be here, you know* (Stalnaker, 1970b). In these cases the horizontal proposition is relevant. Similarly, a sentence of the form *It may be that A* can express that A is consistent with what is presupposed, but it can also express the modal proposition that A is consistent with some suitable chosen set of (physical, logical, ethical ...) law sentences.[40] In the latter case, it is again the horizontal proposition that is relevant.

This horizontal proposition should normally be determined with respect to the worlds compatible with what is presupposed about the subject matter of conversation. This can be shown by another reason why we need to be able to determine the whole propositional concept: *referential disambiguation*. It is obvious that when the speaker uses a demonstrative pronoun like *you* in a sentence like *I will see you at 10 o'clock tomorrow*, he intends to say something about a specific individual, and thus intends to communicate the horizontal proposition. Only because it is sometimes unclear to the hearer what the speaker has intended to say does the diagonal become relevant. Let us assume that I uttered *I will see you at 10 o'clock tomorrow* in a conversation with Antje, Peter and Tim. Suppose that it is common knowledge among us that Antje is going to Berlin this very evening. In that case, even if my pointing has not been very clear, it will be clear to the three hearers what I did not intend to say: namely, that I will see Antje tomorrow at 10 o'clock. Intuitively, this inference follows from the knowledge that if I were talking to Antje, I would be saying something trivially false. However, the inference that I was not talking to Antje cannot be made if the hearers considered only the diagonal expressed by the utterance and the ambiguous pointing. To disambiguate, we need to look at the possible horizontal propositions expressed, where the horizontal propositions expressed are determined with respect to the possibilities of the context.

Determining the whole propositional concept is also needed for another case of ambiguity; the case of *questions*. According to the two-dimensional analyses of questions of Groenendijk & Stokhof (1982), and Lewis (1982), a question denotes its set of true answers. What a true answer is depends, of course, on what the world looks like. Thus, if A and B are the two relative alternatives, the *whether* phrase *whether A or B*, denotes a function from worlds to the proposition expressed by the disjunction of the true alternatives in this world. We can say that for any w, the content of *whether A or B* in $w, [A \veebar B](w)$, is defined as follows: $\{w' \in K |\ (w \models A \text{ and } w' \models A) \text{ or } (w \models B \text{ and } w' \models B)\}$. Thus, if in world w only A is the case, $[A \veebar B](w)$ equals $[A](w)$, and if in w only B is the case $[A \veebar B](w)$ equals $[B](w)$. Let us assume that a context, C, can be represented by a set of worlds. Because some worlds in context C might be A-worlds, and others B-worlds, $A \veebar B$ denotes a set of alternatives with respect to C : $\{[A \veebar B](w) |\ w \in C\}$, the set of possible true answers to the question. But it would be impossible to get such a set

[40] For this reason, Stalnaker (1970b) says that sentences of the form *It may be that A* are *pragmatically ambiguous*.

of alternative propositions if we didn't separate the roles of context and index, since the diagonal proposition expressed by *whether A or B* is just the ordinary disjunction *A or B*. If we didn't look at the horizontal proposition expressed in each world, but only at the diagonal proposition, the question *Will John come?* would state in each world the same trivial proposition: *John will come, or he won't come.*

By separating the roles of reference context and index, we can also account for the two different uses of definite descriptions originally described by Donnellan (1966). Consider the case where the speaker and the hearer see a woman with a man. The man treats the woman kindly, and the speaker and hearer both assume, and know each other to assume, that the man is the woman's husband. However, the assumption is wrong: in fact the man is not the woman's husband. As Donnellan argued, even if their assumption is wrong, we still have the intuition that the speaker has said something true about the person he had in mind in uttering *Her husband is kind to her.*[41] Donnellan (1966) proposes to account for this fact by assuming that definite descriptions can be used in two ways: attributively and referentially. If a speaker uses a description *attributively*, he 'states something about whoever or whatever is the so- and so'; if he uses a description referentially, he 'enable[s] his audience to pick out whom or what he is talking about and states something about that person or thing' (p. 285). It has always been unclear what kind of ambiguity Donnellan was pointing to, but it is generally agreed that we should not think of it as a semantic ambiguity.[42] But then how can we account for the two readings of the above sentence if we assume that definite descriptions are *semantically* unambiguous and always refer to the unique (most salient) individual that satisfies their description in the world under consideration?

Stalnaker (1970b) proposed a straightforward answer to this question: the referential/attributive ambiguity is not semantic, but pragmatic in nature. The referent is either determined by context, or is dependent on the relevant index and determined by the proposition expressed. If a description is used attributively, the rule for determining the referent of the description is part of the horizontal proposition expressed. Accordingly, if in the actual world the man seen with the woman is not her husband, the description *her husband* does not refer to this man in the actual world, but instead to the unique man, if any, who actually is her husband. If the description is used referentially, however, the rule for determining the referent of the description is not part of the horizontal proposition expressed; the expression will instead refer to the individual that is *presupposed* to be the unique (most salient) object that satisfies the descriptive content. If what is presupposed about the denotation of the description *her husband* is the same as what is actually in the denotation of the description, it doesn't matter much whether the description is used attributively or

[41] Of course, when a third person informs the speaker and hearer that they are wrong about this man, that he is not her husband, the speaker can normally no longer use the description *her husband* to refer to this man – if he does use it, this can only be in a special 'ironical' sense.

[42] See Kripke (1977).

referentially. However, there might be a difference if the actual world is not compatible with what is presupposed. If the speaker presupposes of a that he is her husband, then the horizontal proposition expressed by the sentence *Her husband is kind to her* in which the description is used referentially will be that a is kind to her. This proposition might be true in the actual world, although a is not her husband. Thus, by separating the roles of context and index we can account for the intuition that in the circumstances sketched above, the sentence *Her husband is kind to her* have both false and true readings, even though the noun phrase is semantically unambiguous, referring in both cases to the unique (most salient) object that satisfies the description. Again, it is not possible to account for this intuition if one considers only the diagonal proposition expressed by a sentence.

At first sight it might seem that our analysis of referentially-used descriptions has treated them rather differently from the way that we have treated referentially-used pronouns. But this is not really the case. In fact, we can say that if the noun phrase *her husband* is used referentially in the sentence *Her husband is kind to her*, the description will be treated similarly to the pronoun *he* in the discourse *She has a husband. He is kind to her*, where the indefinite *a husband* is used specifically. Because the indefinite is used in this way, it will make a specific object available for reference by a pronoun or a short definite description. This object will be the individual that the speaker has had 'in mind' in his use of the indefinite, the man whom he believes to be her husband. The speaker introduces to the discourse not only an individual but also a *guise*, the subject of the presupposition state corresponding to the discourse referent introduced by the use of the indefinite, under which the individual is represented in the presupposition state. Even if the individual whom the speaker believes to be her husband is actually not her husband, the individual is still represented in the presupposition state as her husband. If the speaker then says *He is kind to her*, he will refer to the individual who was not only the object he had in mind for his use of the indefinite, but also the object that is the *source* of the subject of the presupposition state introduced by his earlier use of the indefinite. Thus, by means of the pronoun, the speaker will refer in the actual world to an individual who is not her husband, just as with Donnellan's case of the speaker using the description referentially. We might say about Donnellan's case, too, that the presupposition state has a subject that is the representation of a particular individual. This individual need not be the same as the value of the subject in all possibilities of the presupposition state, however, it is the *source* of this subject. We associate with this individual the information that we have represented him in our presupposition state under the guise of being her husband. When the speaker uses the description *her husband* referentially, he refers to the same individual that our earlier speaker did when he used the pronoun *he* referentially in the discourse. In both cases he refers to the individual whom we presuppose to be her husband, the source of this relevant subject.

In this section I have followed Donnellan (1966) in claiming that a description like *her husband* might be used referentially, referring to the individual whom the speaker pre-

supposes to be her husband. I have argued that when a description is used referentially, we should analyse it in a way very similarly to when a pronoun is used referentially. In both cases it should refer to the *source* of a relevant subject of the presupposition state. But, then, Donnellan argued that a description might not only be used referentially, but also attributively. I have argued with Stalnaker (1970b) that when a description is used attributively, the rule for determining the referent of the description is part of the (horizontal) proposition expressed. Assuming that pronouns and definite descriptions behave very similar, this suggests that also pronouns might be used attributively. That is, that we associate with a pronoun a description that is interpreted as being used attributively. Indeed, in the following sections I will argue that pronouns might be used attributively, or descriptively, as I will say, and show how this can be implemented within our pragmatic theory.

2.7 Epistemic *might*

Just as in the traditional account of Stalnaker (1978), for *metaphysical* necessity, we check whether the *horizontal* proposition is necessarily true, for *epistemic* possibility, on the other hand, we check whether the *diagonal* proposition is consistent with the context. Indeed we might say that this is exactly what Veltman (1996) proposed in his propositional update semantics. Just as Veltman's propositional update semantics must account for the fact that *It might be that A ... It is not the case that A.* is consistent, but *It is not the case that A ... It might be the case that A.* is not, if we want to account for epistemic *might* within our framework, we must account for the fact that *Someone is P. He might be Q and he is not Q* is consistent, but *Someone is P. He is not Q, and he might be Q* is not. According to the most straightforward analysis, we just assume that *might* is a consistency check with respect to what is presupposed in the possibility as follows:

- $[[\Diamond A]]^{w,c,g,C} = 1 \quad \text{iff} \quad \exists \langle v, c', h, C' \rangle \in C : \ [[A]]^{v,c',h,C'} = 1$

As it happens, something like this was already done by Dekker (1993) within the framework of CCT.[43] But things are somewhat more problematic when we embed epistemic *might* under a quantifier. Let us say that in CCT sentence A is accepted in possibility $\langle w, g \rangle$ w.r.t context S, iff $\langle w, g \rangle \in [[A]](S)$. Now it follows, as Dekker observed himself, that the corresponding analysis within the framework of CCT gives rise to the undesirable result that a sentence like *Someone might have escaped*, represented by a formula like $\exists x \Diamond Ex$, is accepted in $\langle w, g[^x/_d] \rangle$ iff *some* value of x is possibly an E with respect to S. It should however only be accepted in $\langle w, g[^x/_d] \rangle$ iff d itself is possibly an E.[44]

[43] Of course, Dekker (1993) did not account for this in a distributive way, but that is irrelevant for our discussion.

[44] Similarly, $\forall x \diamond Ex$ is predicted to be accepted in any $\langle w, g \rangle$ of S, iff there is the possibility of someone having property E, even if there are some individuals of whom it is known that they don't have property E.

Within our framework we also have a problem with sentences like *Someone might have escaped*, although at first sight this problem seems to be rather different from the problem for CCT discussed above. Our problem is that we are not even able to evaluate a predicate like $\hat{x}\Diamond Ex$, because we don't know how to interpret variable x of the embedded clause. This is true whether we assume that the indefinite is used specifically or unspecifically, and thus whether we have represented the sentence by either $\hat{x}(\Diamond Ex)(\eta r_n \hat{y} Person(y))$, or by $\exists\hat{x}\Diamond Ex$. What we have to do, of course, is to remember which variables we have abstracted over.[45] This can be done when we re-evaluate epistemic *might* as follows:

- $[[\Diamond A]]^{w,c,g,C} = 1$ iff $\exists\langle v, c', h, C'\rangle \in C :\ [[A]]^{v,c',h_g,C'} = 1$
  where $h_g \stackrel{def}{=} \{\langle r, d\rangle \in h|\ r \in DM_L\} \cup \{\langle x, d'\rangle \in g|\ x \in VAR_L\}$

According to this interpretation rule, the variables, but not the discourse referents, occurring in the embedded sentence of $\Diamond A$, evaluated in possibility $\langle w, c, g\rangle$ are interpreted with respect to g. Note that it is now rightly predicted that (i) in case the indefinite was used specifically it is only true when there is a world consistent with what is presupposed where that specific individual that the speaker had in mind has escaped, and (ii) that in case the indefinite was used unspecifically, $\exists\hat{x}\Diamond Ex$ is predicted to be true in possibility $\langle w, c, g\rangle$ just in case there is a particular individual of whom it is consistent with our presupposition that he escaped.[46,47]

Observe that the above interpretation rule predicts that bound variables do not behave in the same way as personal pronouns when they are embedded under the epistemic *might* operator. And indeed, this seems to give the right predictions; when the indefinites in the (a)-sentences are used specifically, the following two pairs of sentences are correctly predicted not to be equivalent:[48]

(27) a. There is someone hiding in the closet. *He* might be the one who did it.

[45]Indeed, something similar has been proposed by Groenendijk et al. (1996) within CCT by changing the interpretation rule for existential quantifiers as follows:

$$[[\exists x A]](S) \quad = \quad \bigcup_{d\in D} [[A]](S[x := d]),\ \text{if } \forall g \in G(S) :\ x \notin dom(g), \text{undefined otherwise}$$

where $G(S)$ is the set of assignments of S. By this new interpretation rule for indefinites, a semantic distinction is made between what Evans called *bound* and *unbound* pronouns. Bound pronouns are interpreted rigidly, as real individuals, whereas unbound pronouns are not.

[46]This is the case because $I_{w,c,g,C}(\hat{x} \diamond Ex) = \{d \in D|\ \exists\langle v, c', h, C'\rangle \in C : [[Ex]]^{v,c',h_{g[x/d]},C'} = 1\}$.

[47]For another, more recent, discussion of this problem and a new one, see Aloni (2001). Her solution does not, however, predict a distinction between on the one hand (27a) and (28a), and (27b) and (28b) on the other.

[48]The sentences are used in Groenendijk et al. (1996) to motivate the interpretation of existential quantification as in the footnote above, thereby giving up the donkey-equivalences of original CCT between $\exists x A \wedge B$ and $\exists x[A \wedge B]$, and between $\exists x A \rightarrow B$ and $\forall x[A \rightarrow B]$, once epistemic *might* is involved. The sentences are attributed to David Beaver.

b. There is someone hiding in the closet *who* might be the one who did it.

(28) a. If there is someone hiding in the closet, *he* might be the one who did it.

b. Anyone who is hiding in the closet might be the one who did it.

The non-equivalences can be illustrated by the following model:

$$W = \{w, w'\}, D = \{d, d'\}, I_w(Hide) = \{d\}, I_{w'}(Hide) = \{d'\}, I_w(Did\,it) = I_{w'}(Did\,it) = \{d\}$$

Given that we presuppose that either w or w' is the actual world, after interpreting (27b) we want to end up in the information state that we are in w – that is, that d is hiding in the closet. After updating our presupposition state with (27a), however, we still don't know in what world we are, and thus who did it. The same holds for (28b) and (28a), respectively.

Note that according to our interpretation rule it follows that epistemic *might* quantifies over world/reference context/assignment triples. It might be questioned, however, whether epistemic *might* really quantifies over such fine-grained entities. Would the speaker of *It might be that A* ever express more than only his incomplete knowledge about the *subject matter* of the discourse? If not, then it is natural to let *might* quantify only over worlds, in the following way:

- $[[\Diamond A]]^{w,c,g,C} = 1$ iff $\exists \langle v, c', h, C' \rangle \in C : [[A]]^{v,c',g,C'} = 1$[49]

According to this interpretation rule the embedded sentence is not interpreted with respect to a shifted assignment function, so we don't have any problem with sentences where epistemic *might* is embedded under a quantifier. But two other problems do now arise: First, we can no longer make a distinction between '$\Diamond x = y$' and '$x = y$' if either x or y is a discourse referent; and second, the (a) and (b) sentences mentioned above are predicted to be equivalent when the indefinites in the (a) sentences are used specifically.[50] Is there a natural way to let the (a) and (b) sentences not be equivalent, and still assume that epistemic *might* quantifies only over worlds? There is, if we make use of *descriptive* pronouns.

We might say that the first analysis of epistemic *might*, where it also quantifies over assignment functions, comes down to this: if the speaker says *Someone committed the murder. It might be the butler*, the speaker has used the indefinite referentially and thus has a specific individual in mind, but doesn't know whether or not he is the butler. This might well be the right way to analyse such discourses, however, there is an alternative that seems at least as natural as the one above. According to this alternative analysis, the speaker has no specific individual in mind by his use of the indefinite; and the pronoun *it*

[49] This interpretation rule is similar to the one proposed by van Eijck & Ceparello (1994).

[50] At least, if it is the truth of the discourse that counts for (27a).

is not used as a referential pronoun but as a *descriptive pronoun* instead, going proxy for *the one who committed the murder*. Notice that once we assume that the pronoun is used descriptively, we can assume that a speaker who asserts *It might be that A* expresses only his incomplete knowledge about the *subject matter* of the discourse, and thus that *might* quantifies only over worlds.

Now we can also explain why (27a) and (28a) are not equivalent to (27b) and (28b), respectively, without assuming that epistemic *might* also quantifies over assignment functions. The explanation now is that the personal pronouns in the (a) sentences are used *descriptively*, going proxy for *the person who is hiding in the closet*: and that (ii) the relative pronouns in the (b) sentences cannot be treated as descriptive pronouns, because they are c-commanded by a quantifier, and should thus simply be treated as bound variables. As can be easily checked, when the pronouns in the (a) sentences are treated as E-type pronouns, and those in the (b) sentences as bound variables the (a) sentences are predicted to be less informative than the (b) sentences.[51]

In this section we have seen that it might be possible to let epistemic *might* quantify only over worlds if we assume that pronouns can sometimes be used descriptively. Until now we have considered only one piece of 'evidence' for the existence of descriptive pronouns, and this evidence is admittedly not very clear. Fortunately, there is more compelling evidence for the existence of descriptive pronouns. It is to this evidence that we now turn.

2.8 Descriptive pronouns

One of the main motivations in section 2.3 for treating at least some uses of pronouns as being referential was the phenomenon of *pronominal contradiction*. A pronoun sometimes refers back to the speaker's referent of the antecedent. But as Kripke (1977) noted, the same phenomenon shows that sometimes pronouns also refer back to the *semantic referent* of the antecedent. If A says *Her husband is kind to her*, B can react by saying *No, he isn't. The man you are referring to isn't her husband*. The pronoun *he* refers now to the semantic referent of its antecedent, the actual husband of her. Thus, pronouns can be used in two

[51] This seems like a nice picture. Still, Groenendijk et al. (1996) have given an example where the extra fine-grainedness of subjects seems to be needed anyway. Consider the following mathematical example:

(i) $\exists x[x^2 = 4] \wedge \Diamond x = 2 \wedge \Diamond x = -2$

They claim that this example shows why we need to quantify over assignments, because the first claim doesn't give us new information about the world; in every world 4 is a square of 2 and -2. But I'm not completely convinced by this example. I find the discourse "There is a number whose square is 4. It might be 2, and it might be -2" very unnatural. I find it much more natural for a speaker to say something like "There is a number on this card whose square is 4. The number might be 2, and it might be -2. Can you guess which number it is?" But in this case the definites *the number* and *it* can simply stand for the description *the number on this card*, so that this case would not constitute a counterexample to the second way of interpreting epistemic *might*.

ways: they either pick up a previous semantic reference or a previous speaker's reference.

In section 2.3, I argued that pronouns are normally used referentially, referring to the unique individual that is the speaker's referent of the antecedent of the pronoun. According to the motivation I gave for our analysis it can be explained why an indefinite used under the scope of two negations can normally not be taken as syntactic antecedent for pronouns used in the 'main' context. The reason is that the indefinites occurring in those positions normally have no actual speaker's referent, something that is required if the pronoun is used referentially. In a similar way it can also be explained why an indefinite used in one disjunct can normally not be taken up by a pronoun occurring in the other disjunct. In original CCT as given in Appendix B, such constraints are given by *syntactic* means. Unfortunately, there are well-known counterexamples to these constraints on anaphoric binding: pronouns can sometimes take indefinites as their syntactic antecedent even though the anaphoric island constraints predicted by DRT/FCS are violated. It has been noted that these cases can be accounted for by assuming that the pronouns involved are interpreted as E-type pronouns. That is what I will argue for, too– [52] although my proposal of this division of labour, unlike similar ones, will give a natural motivation for this division of labour, together with a formal account of E-type pronouns within (my version of) dynamic semantics. The motivation for this division of labour is the following (which I gave above): What a speaker refers to with his use of a pronoun depends on his intentions; pronouns normally are used referentially, and it is specific intentions that count for them. Sometimes, however, pronouns can be used descriptively, and here it is only the general intentions that count. In the latter cases, a pronoun refers to (all of) the (unique) individual(s), if there is (are) any, that satisfies the description associated with the pronoun that is recoverable from the antecedent clause. Note that we assumed that pronouns can be used descriptively, there is no longer any reason to expect that they cannot escape the DRT/FCS anaphoric island constraints. Instead, the constraints on the appropriate use of descriptive pronouns can be given in terms of what is presupposed to be true.[53]

In this section I will concentrate on only one kind of case in which the constraints on anaphoric binding in standard dynamic semantics are too rigid. I will focus on a paper by Krahmer & Muskens (1995), who make the implicit claim that we don't have to rely on the existence of descriptive pronouns to account for some apparent counterexamples to CCT.[54] I will argue, however, that their proposal leaves something unexplained, and that this something can be accounted for naturally by the E-type approach. Later, I will account for the existence of descriptive pronouns in (my version of) dynamic semantics.

Consider the following sentences:

(29) Either John does not own *a donkey*, or he keeps *it* very quiet. (Evans, 1977)

[52] Although the most convincing case for the existence of E-type pronouns involves plural pronouns, I will limit myself in this section to singular pronouns.

[53] Perhaps after accommodation.

[54] Somewhat similar proposals are made by Groenendijk & Stokhof (1990), and Dekker (1993).

(30) Either there is *no bathroom* in the house, or *it*'s in a funny place. (Roberts, 1989)

(31) It is not true that John didn't bring *an umbrella*. *It* was purple and *it* was standing in the hallway. (Muskens & Krahmer, 1995)

It is well known that standard CCT has problems with such sentences. The reason is that in CCT negation is treated as a plug with respect to anaphoric binding. Note that on an E-type account negation does not have this property. Proponents of the standard dynamic account argue that negation *should* be treated as a plug, because this is the most natural way to account for the unacceptability of (32):

(32) There is *no guest* at this wedding. *He* is standing right behind you.

In other words, CCT can account for the unacceptability of (32) by *syntactic* means: an object 'introduced' under the scope of a negation cannot be picked up by anaphoric means in further discourse. But the E-type approach, of course, has no problem accounting for the unacceptability of (32) either. The sequence (32) is unacceptable, not for syntactic but for *semantic* reasons. The context resulting after the interpretation of the first sentence of (32) contains no world in which there is a guest at this wedding. If the pronoun *he* of the second sentence stood for *the guest at this wedding*, then the second sentence would be trivially false. That's the reason why (32) is out. This reasoning about (32) seems natural, I would say. And does the acceptability of the sentences (29), (30) and (31) not justify this reasoning also?

Not so, say Krahmer & Muskens (1995). Negation is a syntactic plug with respect to anaphoric binding, and the reason why sentences (29) - (31) are acceptable is that a double negation is a plug unplugged. A clause of the form '$\neg\neg A$' is not only truth-conditionally but also *dynamically* equivalent to 'A'. They account for this claimed equivalence in a way that is not completely ad hoc by using techniques from partial logic.[55]

I have some worries about their approach, however. First, intuitively there seems to be no difference between (29) and a sentence like (33):

(33) a. It is possible that John does not own *a donkey*,

b. but it is also possible that he keeps *it* very quiet.

It would be good if both could be handled by the same mechanism. But it is rather doubtful that this mechanism could be that $\neg\neg A$ is equivalent to A. Second, if an indefinite is used under the scope of two negations, it seems that a singular pronoun can take it as syntactic antecedent only if there is only one object (in each of the relevant worlds) that could be the referent of the indefinite. For (29) and (30), for instance, the uses of the pronoun

[55] Groenendijk & Stokhof (1990) and Dekker (1993) reach a similar result by using lifting, instead of partiality.

it in the second disjuncts can only pick up the *unique* donkey that John owns and the *unique* bathroom in the house, respectively. If it is presupposed that John possibly owns more donkeys, and if there are perhaps more bathrooms in the house, the uses of *it* in the respective second disjuncts would be, I think, inappropriate.

My worries, however, are not limited to disjunctions. I think that if an indefinite is used under the scope of two negations, a singular pronoun that is not under the scope of these negations can *never* take the indefinite as a syntactic antecedent if there are more objects in one of the worlds that the indefinite could have referred to. It seems that Krahmer & Muskens agree. Discussing the contrast in acceptability between (34) and (35),

(34) It is not true that there is no guest at this wedding.
?*He* is standing right behind you.

(35) It is not true that there is no bride at this wedding.
She is standing right behind you.

they say that the distinction is due to a uniqueness effect.

> Given some highly unlikely context in which it is understood between speaker and hearer that at most one guest can be present at this particular wedding (34) would be fine. We feel that it is precisely the unlikelihood of such a context which explains the markedness of (34). (Krahmer & Muskens, 1995, p. 359)

I completely agree. But then they make the following claim about these problematic cases:

> Since such apparent counterexamples on closer examination turn out to be no counterexamples at all, it seems we can take it as a general rule that as far as truth conditions and the possibility of anaphora are concerned double negations in standard English behave as if no negation at all were present. (Krahmer & Muskens, 1995, p. 359)

I'm afraid that I don't understand this. That you can explain why a counterexample to your approach *is* a counterexample doesn't mean that on closer examination it 'turns out to be no counterexample at all'.[56]

I propose to take the counterexample seriously. Although original CCT predicts that an indefinite will not be accessible as a syntactic antecedent to a singular pronoun, the indefinite does turn out to be accessible when speaker and hearer both presuppose that there is exactly one object that the indefinite can refer to.

[56]It is sometimes assumed that we can account for bathroom sentences by representing sentences of the form '$\neg P$ or Q' by something like '$\neg P \vee (P \wedge Q)$'. But this gives rise to the same problem the Krahmer & Muskens approach does.

I wish to propose that the division of labour between *referential* and *descriptive* pronouns should be taken seriously in the following way: Where referential pronouns take specifically used indefinites as antecedents, E-type- (or descriptive) pronouns can take indefinites as antecedents only when the speaker has no specific individual in mind.[57] Moreover, a *singular* descriptive pronoun can be used appropriately only if the associated description is presupposed to have only a *unique* instantiation.[58] The dependencies of different kinds of pronouns on different kinds of indefinites, and the uniqueness condition can be illustrated by the contrast in acceptability between the following two sequences:

(36) a. I saw a French movie yesterday. *It* was dreadful.

b. I have seen a French western movie before. *? *It* was dreadful.

As observed by Kálmán and Rádai (1998), the pronoun *it* in (36b) can refer only to a French western I saw under the assumption that I saw exactly one French western in my life, which is not true for (36a). They claim that this distinction is due to the different uses of the indefinite noun phrases: In (36a) the indefinite introduces a discourse referent, while in (36b) it does not, because there the indefinite is used *existentially*. I completely agree with what they say, if I understand their claim as follows. When they say that an indefinite introduces a discourse referent they mean that the indefinite is used specifically. When the indefinite is used existentially, we still want to introduce a variable or discourse referent to the discourse, because we can refer back to this indefinite. In this case, however, the variable will denote a property that can determine the referent of a descriptive pronoun.

Further motivation for this uniqueness condition is the contrast observed by Partee (1972) between the following sentences:

(37) a. John was looking for *the man who murdered Smith*, and Bill was looking for *him* too, and

[57] But at least for *definite antecedents*, this doesn't seem to be quite right. Consider Kripke's (1977) pronominal contradiction example again. If *A* says *Her husband is kind to her*, he has a specific individual in mind, still *B* might react by saying *No he isn't. The man you are referring to isn't her husband*, where the pronoun is used descriptively. So, it seems that even when the speaker has a specific individual in mind for his use of an (in)definite, another speaker can still use a pronoun *descriptively* that takes this (in)definite as an antecedent, and thereby does not refer to the same individual as the first speaker has had in mind for his use of the antecedent.

[58] There is at least one singular pronoun that intuitively picks up more objects, although the pronoun itself *refers* only to one object. The pronoun I have in mind is *one*. As is commonly assumed, this singular pronoun takes a noun as a syntactic antecedent. Consider Partee's (1972) *John lost a black pen yesterday, and Bill found a grey one today*. What is relevant here is that this pronoun is a special kind of E-type pronoun in that it takes up properties and violates the CCT constraints on anaphoric binding in exactly the same way as other descriptive pronouns do. Note, for instance, that the following variant of (34) is perfectly fine:

(i) It is not true that there is no guest at this wedding. *One* is standing right behind you.

b. John was looking for *a gold watch*, and Bill was looking for *it* too.

The pronoun *him* in (37a) can be used when the speaker has no particular man in mind, but the pronoun *it* in (37b) cannot be used when there is no particular gold watch in mind. The source of this contrast, according to Partee, is that it can be presupposed that for (37a) but not (37b) that the uniqueness constraint is satisfied, i.e. that there is exactly one murderer of Smith but not one gold watch.

In order to analyse E-type pronouns within our account, then, we have to implement the following ideas. The first, already implemented, is that indefinites can be used in two ways: specifically and existentially. The second is that referentially-used pronouns can take only specifically used indefinites as antecedents. The third is that a singular pronoun may take an existentially used indefinite as its antecedent, but only if it is presupposed that there exists exactly one object that could be denoted by the property associated with the antecedent sentence.[59] The last, which follows from the claim that descriptive pronouns do not have to obey the accessibility constraints, is that negations should not be treated as absolute syntactic plugs with respect to anaphoric binding.[60,61]

So, what we have to do is (i) add the interpretation rules for the *existential* quantifier and show how both they and definites can introduce properties to the discourse; (ii) show how *descriptive pronouns* that take up these properties can be interpreted; and (iii) change the earlier definition of $Upd(\neg A, \langle w, c, g\rangle)$.

In the new formalisation we make use of the set G^* of partial assignments, where G^* is $\{\mathbf{D}^X |\ X \subseteq VAR_L\} \cup \{[W \rightarrow \wp(D)]^X |\ X \subseteq DR_L\}$. Hence, technically, discourse referents are always assigned properties – although some of these properties represent ordinary objects. The set $\mathbf{D}$ of rigid 'concepts' is defined by:

$$\mathbf{D} \quad \stackrel{def}{=} \quad \{\lambda w.\{d\} |\ d \in D\}$$

Specifically-used indefinites will introduce discourse referents into the discourse that are assigned to rigid 'concepts',[62] while discourse referents introduced by existential quantifiers or descriptions can be assigned any kind of property. Discourse referents representing singular pronouns are now evaluated as follows:

$[[r]]^{w,c,g} = d$, if $g(r)(w) = \{d\}$, and $\forall\langle v, c', h\rangle \in S : h(r)(v)$ is a singleton set
undefined otherwise [63, 64]

[59] For simplicity I will assume that descriptive pronouns can take only non-specifically-used (in)definites as antecedents.

[60] I will do the same later for other constructions that are normally treated as plugs with respect to anaphoric binding.

[61] I am, of course, not the first to analyse E-type pronouns as pronouns that pick up contextually-given properties; Cooper (1979) did the same. What distinguishes my approach from Cooper's is that (i) I do not assume that all discourse anaphora should be treated as E-type pronouns, and (ii) I analyse E-type pronouns within a theory of context change.

[62] Thus, C is from now on a set of functions from indices to $\mathbf{D} \cup \{*\}$.

Thus, a singular pronoun will always refer to the *unique* instantiation in the relevant world of the property associated with the variable. We have already seen in section 2.3 how we can straightforwardly determine the truth value of an existential sentence. What we have to do now is to explain how existential sentences can introduce properties into the discourse and how negation can be treated in such a way that it is no longer an absolute plug with respect to anaphoric binding. Before we do this, however, we have to say under which discourse referent an existential sentence introduces a property. This can be done by representing existential sentences by $\exists r\hat{x}A$ rather than by formulae like $\exists\hat{x}A$, as we have been doing. The dynamics can now be determined by the following definitions of $Upd(\exists r\hat{x}A, \langle w, c, g\rangle)$ and $Upd(\neg A, \langle w, c, g\rangle)$ (ignoring the change in presupposition state and with k/g indicating the result of subtracting g from k):

- $Upd(\exists r\hat{x}A, \langle w, c, g\rangle) \quad = \quad Upd(A, \langle w, c, g[^{r}/_{|\hat{x}A|^{c}_{g}}]\rangle)$
- $Upd(\neg A, \langle w, c, g\rangle) \quad = \quad g \cup \{\langle r, o\rangle \in k/g |\ \exists c' :\ k\ = Upd(A, \langle w, c', g\rangle)\ \&$
  $\forall c'' :\ Upd(A, \langle w, c'', g\rangle)(r) = k(r) = o\}$

The property $|P|^{c}_{g}$ introduced by existentially-used indefinites is that function $f :\ W \rightarrow \wp(D)$ such that for any $w \in W$:

$$f(w) \quad = \quad I_{w,c,g}(P)^{65}$$

The dynamics of existentially used sentences is rather straightforward: we simply introduce the properties associated with their descriptive contents into the discourse. The dynamics of negation have had to be somewhat more involved to account for the intuition that (i) indefinites under the scope of the negation will (normally) not introduce specific individuals and (ii) properties can be introduced by such indefinites, but (iii) indefinites under the scope of a negation do not introduce properties to the main level that are *dependent* on terms standing in monotone-decreasing positions whose referents are not yet established. For instance, I don't want to introduce properties corresponding to *a woman* in *If a man buys a flower, he gives it to a woman*, because the property introduced by this indefinite in the consequent depends on the referents of *a man* and *a flower* in the antecedent of the conditional.[66] The interpretation rule has the result that the properties introduced by $\neg A$ are those introduced by subformulae $\exists rP$ of A that introduce only a single property to the main context.

[64] As usual, the undefinedness of terms carries over to the undefinedness of formulae in the obvious way, which I leave to the reader to determine.

[64] Note that although I assume that pronouns can be used in two different ways, technically pronouns will not be ambiguous: they are always represented by discourse referents.

[65] Or perhaps the reference context should also be shifted, such that $f(w)$ should denote the following set: $\{d \in D :\ \exists c' \in C\ \&\ [[A]]^{w,c',g[^{x}/_{\mathfrak{a}}]} = 1\}$, if P is of the form $\hat{x}A$.

[66] In section 2.10, however, I will allow these indefinites to introduce functions from worlds and individuals to properties.

Let's now discuss the bathroom sentence (30), represented by $\neg\exists r\hat{y}Py \vee Qr$. Suppose that for each world $\langle v, c'\rangle$ consistent with what is presupposed it holds that $kard(I_v(P)) \leq 1$; then the following results:

$$[[\neg\exists r\hat{y}Py \vee Qr]]^{w,c,g} = 1 \text{ iff } [[\neg\exists r\hat{y}Py]]^{w,c,g} = 1 \text{ or } [[Qr]]^{w,c,Upd(\neg\exists r\hat{y}Py,\langle w,c,g\rangle)} = 1$$

$$\text{iff } [[\neg\exists r\hat{y}Py]]^{w,c,g} = 1 \text{ or } [[Qr]]^{w,c,g[^{r}/_{|\hat{y}Py|^{c}_{g}}]} = 1$$

$$\text{iff } I_w(P) = \emptyset \text{ or } (kard(I_w(P)) = 1 \text{ and } I_w(P) \cap I_w(Q) \neq \emptyset)$$

If we assume that in every world consistent with what is presupposed there is at most one P, the singular pronoun represented as r in Qr can be interpreted.

My analysis predicts that the sentence *Either there is no bathroom in the house or <u>it</u> is in another place* introduces the property *bathroom in the house* into the discourse, and that this property can in principle be taken up by a pronoun in a subsequent sentence. In practice, however, this will typically not happen. Is this problematic for my analysis? Given the definedness conditions on discourse referents (and atomic formulae), I don't think so. My theory predicts that one can appropriately use a singular E-type pronoun interpreted as *the P* only if in all worlds consistent with what is presupposed there exists exactly one P. But given that we have used a disjunction, and that the existence of a P is precisely what is at issue here, it will typically not be true in such circumstances that there is one and only one P.[67]

Note that E-type pronouns in sentences of the form *Either <u>she</u> has no husband, or he is not here* are generally speaking interpretable if the pronoun *she* is interpretable. This is in general the case. As it happens, this is required to account for the fact that E-type pronouns can be *relational* and *indexical* (sec. Neale, 1990):

(38) *Smith's murderer* is insane. *He* should be jailed for life.

(39) The one who wins *this game* will be lucky. *He* will get all the money.[68]

Up to now I have concentrated on indefinite antecedents. But of course, you can also refer back with a singular pronoun to a *definite description* where standard dynamic semantics predicts this to be impossible. Just like indefinites, they introduce a property. What is special about non-anaphorically used (singular) definite descriptions is that in

[67] But sometimes there will be, as in *Either there is no bathroom in the house, or it is in a funny place. In any case, it is not on the ground flour.* Note that for this example it is crucial that the pronoun in the second sentence be a descriptive pronoun and that the description have a scope smaller than the negation.

[68] Neale argues that many incomplete definite descriptions can be completed by purely referential or indexical material. I agree that such a claim is natural for descriptions like *the mayor* or *the murderer*. But I don't think that descriptions should normally be treated as Russellian descriptions (see also Evans, 1982, pp. 324-325).

every world of the context resulting from the interpretation of this description there will be one object only that satisfies the description. This has two consequences. One is that the concept introduced can be restricted to the noun phrase itself. The other is that one is guaranteed to be able to refer back to such a description when considering a world in the same context as that in which the description has been used. The dynamics of (attributively-used) definite descriptions is rather straightforward: the iota term ιrP simply introduces into the discourse the property corresponding to P.

Notice that the theory predicts that singular pronouns can sometimes pick up definite descriptions introduced into positions predicted to be inaccessible on the standard CCT account. Here are some examples that suggest that singular pronouns can indeed do so:[69]

(40) If John makes coffee, *his wife* will be happy. *She* is a nice person. (after V.d. Sandt, 1992)

(41) If all countries have presidents, *the president of France* probably regards himself as their cultural leader. *He* is such a pompous ass. (Geurts, 1995)

These examples are supposed to show that the definites in the consequents should have scope over the whole conditional.[70] How else could we interpret the unbound pronoun in the second sentence? The argument is a forceful one if we adopt the assumption that negations are absolute plugs with respect to anaphoric binding. But on our account, the argument loses its force. On the assumption that John has only one wife, a discourse referent will be introduced into the 'main' context in a systematic, compositional, and non-representational way that can be picked up by a singular pronoun. Note that if we substituted *a man* for *John* in the antecedent of (40), the description *his wife* could not be anaphorically picked up by a subsequent sentence, which is indeed what I predict.[71]

Because the interpretation of a proper name depends on worlds and not on reference contexts, we can also account for the fact that proper names in positions that, on a standard dynamic semantics account, make introduced discourse referents inaccessible can always be picked up in the ungoing conversation. In contrast to other proposals,[72] this one doesn't require a special proper name rule to account for this.

Evans (1977) argued that E-type pronouns are *referential* expressions, referring to the individuals satisfying its descriptive content. We have chosen to follow Neale (1990), however, in interpreting E-type pronouns as *descriptive* pronouns. This descriptive analysis has two advantages: it allows us to treat so-called *concept anaphora* as E-type pronouns; and it allows us to make the interpretation of the pronoun dependent on the *scope* of the discourse referent by which the pronoun is represented.

[69] On the assumption that the definite descriptions are not interpreted referentially.

[70] Although this is perhaps not the way van der Sandt and Geurts would phrase it.

[71] Van der Sandt predicts this too, but for purely *syntactic* reasons: the definite description *his wife* cannot take scope over the conditional because if it were, the pronoun *his* would not be interpreted.

[72] See Kamp & Reyle (1993), for instance.

Note that our interpretation rules predict the possibility of a pronoun picking up a description interpreted in a world, or a more complex index, that is not an element of the set of indices of the context resulting from the interpretation of the indefinite. This is good news if we want to account for the following examples, in which so-called concept anaphora is involved:

(42) *My home* once was in Maryland, but now *it*'s in Los Angeles. (Partee, 1972)

(43) a. Senator Green believed that he had nominated *the winner of the election*,

b. but Senator White believed that she had nominated *him*. (Partee, 1972)[73]

(44) This year *the president* is a Republican. Next year *he* will be a Democrat. (Evans, 1977)

(45) John believes that *the winner of the game* needs to play well, while Mary believes that *he* must just be lucky.

Note that because on our account variables can stand in scope relations to other constructions, the interpretation of the pronoun might depend on the position it is interpreted in. Of course, position is irrelevant for referential pronouns;[74] but it *is* relevant for descriptive pronouns. For instance, the denotation of the E-type pronoun *he* in the following example depends on whether the pronoun has wide, narrow, or intermediate scope:

(46) The mayor is a democrat. John *thinks* that *next year* he'll be a republican. (Neale, 1990, p. 214)

That is, the second sentence in this example can have the following three representations:[75] $\hat{y}Bel(j, NY(Republic(y))(r)$; $Bel(j, NY(Republic(r)))$; and $Bel(j, \hat{y}NY(Republic(y))(r))$.

2.9 Plurals, quantifiers, and functional pronouns

A speaker can use a singular pronoun appropriately when in every possibility of the context there is an object available for reference to which this pronoun refers. There are two ways in which an object could become salient or available for reference. Either because it is observable for both speaker and hearer or because (normally) the speaker made it salient by making it clear, for instance by using an indefinite description, that he has a specific

[73]Partee (1972) notes that (43) is ambiguous: the senators argue either about who has nominated a certain person, or about who the winner of the election will be, the one nominated by Green or the one nominated by White. She concludes that a sentence like (43) "constitutes a real problem for any attempt to find a uniform basis for the pronoun-antecedent relationship" (p. 425).

[74]More specifically, for referential *singular* pronouns. However, as we shall see below, scope can be relevant for plural pronouns, even when these pronouns are used referentially.

[75]Where NY stands for next year.

object 'in mind'. Obviously, not only single objects but also sets of objects can be available for reference in these ways. There might, for instance, be obvious criteria to select subsets of observable entities in the environment that the speaker and hearer share; and the speaker can make a set of individuals salient by using a plural indefinite. We can refer to such sets by plural pronouns. For the sets available for reference by observable criteria we have deictic and demonstrative uses of the plural pronoun *they*. We also have the anaphoric, but non-E type, use of *they*, as in the example below, repeated from (18):

(47) Yesterday, John met some girls. *They* invited him to their place.

Plural pronouns can refer back to salient sets; but other nominal expressions can also perform this function. If a speaker says *Everybody had a good time*, he is probably not claiming that everybody in the universe had a good time, but is restricting his domain of quantification to a certain set of individuals. For the assertion to be understandable for the hearer, this set of individuals must be salient somehow. It's natural to assume that such a quantifier can restrict its domain of quantification by the same sets that are available for reference for plural pronouns. This suggests that the interpretation of quantified noun phrases in a possibility of the context depends in the same way as plural pronouns on the reference context of that possibility. The domain of quantification can be determined either by deictic or by anaphoric means. Quantifiers are not two- but *three*-place relations. Indeed, this has already been proposed by Westerståhl (1984) and van Deemter (1991). Because in certain situations a sentence like *All were happy* makes sense, a salient context set is sometimes needed to interpret a sentence in which a certain anaphoric quantifier (determiner) occurs. Westerståhl even shows that we need several salient context sets distinct from the domain of discourse to give a reasonable interpretation of sentences containing more than one quantifier. Consider the following example of Westerståhl's:

(48) a. The English love to write letters.

b. *Most* children have *several* pen pals in many countries.

To make sense of this discourse, we have to assume that the domain of discourse contains both English and non-English children: although *most* is restricted to the set of English children, *several* is not.

Quantified phrases can thus be anaphoric, in that their interpretation depends on some salient context set. But they can also *introduce* context sets into the discourse, which we can refer back to by means of other anaphoric quantifiers or plural pronouns (see e.g. Kamp & Reyle (1993), van den Berg (1996), van der Does (1994), and Fernando (1994)). This can be illustrated by the following sentences:

(49) a. Most Dutch farmers have financial problems.

b. Most older farmers think about quitting.

(50) a. Fred bought most donkeys.

b. Then he sold *them*.

The noun phrase *most older farmers* in (49b) is most naturally interpreted partitively as denoting most older Dutch farmers that have financial problems; and the pronoun *them* in (50b) is most naturally interpreted as denoting all of the donkeys that Fred bought. Thus, it seems natural to represent a quantified sentence of the form '[Det A] B' or '[ADV A] B' in general by '$Q^x_y(A, B)$' where Q is the relevant determiner or adverb; x the variable that represents the set that the quantified phrase anaphorically referred back to; and y the variable that represents the introduced set, or better, property.

Plural pronouns can refer back to properties introduced by quantified phrases. As we have already seen, the abstraction operator allows us to have terms – in particular, discourse referents that represent pronouns – standing in non-trivial *scope* relations with other constructions. Just as with singular E-type pronouns, it is also important that plural E-type pronouns can be interpreted descriptively. Together with the possibility that the variable representing the pronoun can stand in various scope relations with other constructions, this descriptive interpretation allows the plural pronoun *they* in a discourse like (51) to be given not only a *de re* but also a *de dicto* analysis, each corresponding to a possible reading of the discourse

(51) Most friends of Sue will marry a Swede. Sue believes *they* will be happy.

The scope of pronouns is important not just in intensional contexts. As we have already seen, scope is relevant for singular pronouns only when the pronoun is used descriptively. For plural pronouns, however, scope can even be important in case the pronoun is used referentially. A sentence like *They walk* is true just in case everybody in the relevant context set walks. Similarly, on the most natural reading of the sentence *They don't walk*, the sentence seems to be true just in case none of the individuals in the relevant context set walks. How can we account for this reading? Dekker (1994) suggests that we need to resort to truth-conditionally partial semantics to give a semantic account of plurals. But we don't need partial semantics for this purpose. With the use of our abstraction operator, we can simply say that the pronoun *they* has wide scope with respect to the negation. Thus, the scope of a quantified pronoun can matter even in an extensional context.

Until now I have assumed that quantifiers and plural pronouns anaphorically refer back to a salient context set and that quantifiers simply introduce *sets* (or properties) to the discourse. What makes the dynamics of (adverbial) quantifiers and plural pronouns so complex, however, is that the interpretation of such quantifiers or pronouns which can anaphorically refer back to other quantifiers, can be dependent on the interpretation of still other quantifiers, as in (52)

(52) *Every man* loves *a woman*. *They* prove this by giving *them* flowers.

On the most natural, *distributive*, reading,[76] the interpretation of *them* depends on the interpretation of *they*. As a result, on at least one reading of the second sentence every man who loves a woman proves this love to *the woman he loves* by giving her flowers. The problem now is to interpret quantifiers and pronouns in such a way that this dependence can be accounted for.

This dependence occurs not only with plural but also with singular pronouns, as can be shown by the following examples:

(53) Most summers John rents *a car* to go to France. He usually takes *it* on a ferry.

(54) Every man lost *a pen*, and some man found *it*.

(55) *Every player* chooses *a pawn*. *He* puts *it* on square one. (Roberts, 1989)

In all of the above examples, the interpretation of *it* in the second sentences depends on something: on the relevant summer in (53), on the relevant man in (54), and on the relevant player (the referent of *he*) in (55). Note that in all these examples, the pronoun seems to go proxy for a *possesive*.

Although some analyses of 'dependent pronouns'[77] have avoided treating them as *functions* going proxy for *descriptive* phrases I think that it is natural to give a treatment of these pronouns similar to that given for descriptive pronouns. I have two reasons for thinking this. First, in the examples above, just as with descriptive pronouns, there seems to be a notion of *uniqueness* involved. If, for instance, some players are presupposed to choose more than one pawn, the use of the singular pronoun *it* in (55) does not seem appropriate. Second, an analysis of dependent pronouns as functional pronouns can also be extended to another kind of pronoun, which exhibits a dependency very similar to the concept anaphora discussed earlier. This, of course, is the famous paycheque pronoun described by Karttunen (1969), which is exemplified in the sentence below:

(56) The man who gave *his paycheque* to his wife was wiser than the man who gave *it* to his mistress.

Note that with the machinery introduced in section 2.8, I still cannot account for such paycheque examples, although the E-type approach is usually assumed to be appropriate here, too.[78] The reason I cannot yet account for these should be obvious: the functions I introduce are functions from worlds to sets of individuals; in general, however, I should introduce functions from world-*sequence* pairs to sets of individuals. In section 2.8 I accounted for the case in which the sequence is the empty sequence; in that case the pronoun

[76] I will ignore here the collective reading. For a different, and much more elaborate discussion of the interpretation of plural pronouns in dynamic semantics, see van den Berg (1996).

[77] See for example van den Berg (1996), Fernando (1994), and van Rooy (1998).

[78] See Chierchia (1996), for instance.

goes proxy for a definite description recoverable from the antecedent clause. In this section I want to account for the case in which the sequence consists of one or more individuals. In the case of one individual, we might say that the pronoun goes proxy for a *possessive* recoverable from the antecedent clause.

In order to analyse paycheque pronouns in the way suggested above, I will say that possessives like *the paycheque of* are represented by something like $\hat{x}(\iota r[Paycheque - of(r,x)])$; and that pronouns can be represented by complexes like '$r(t)$', where in each world, r denotes a function from individuals to sets of individuals and t denotes an individual. We need to know two things: (i) how pronouns represented as complexes like '$r(t)$' should be interpreted, and (ii) how our more complex functions should be introduced.

The first question can be answered straightforwardly: a functional pronoun represented as $r(t_1,...,t_n)$ is interpreted in $\langle w,c,g\rangle$ as $[[r]]^{w,c,g}([[t_1]]^{w,c,g},...,[[t_n]]^{w,c,g})$. The answer to the second question is somewhat more involved. I will propose that for the dynamics of definites, existentially-used indefinites, and (adverbially) quantified phrases our possibilities have to be enriched by a sequence of variables. The idea is that the function introduced, for instance, by the definite term ιrP with respect to possibility $\langle w,c,\vec{x},g\rangle$ is a function not from worlds to the P's in that world, but rather from worlds and a sequence of n individuals to a set of individuals. It might be easiest, first of all, to give the function denoted by the object $|P|_g^{z,c}\vec{x}$, where each x_i is either a variable or a discourse referent.[79] This object will denote the function $f:\ (W\times D^n)\rightarrow \wp(D)$ such that for any $w\in W$ and $\vec{d}\in D^n$:

$$f(\langle w,\vec{d}\rangle)\ =\ \{e\in g(z)(w)|\ e\in I_{w,c,g[\vec{x}/\vec{d}]}(P)\}$$

Now we need to know how such a function is introduced and, in particular, how the sequence of variables that determines the arity of the function is determined. This is accounted for by the following dynamic interpretation rules (where t is a plural discourse marker, s a sequence of variables, and if $s=\ \langle x_1,...,x_n\rangle$, then $s[y]\ =\ \langle x_1,...,x_n,y\rangle$):[80]

- $Upd(Q^x_{y_t}(A,B),\langle w,c,s,g\rangle)\ =\ Upd(A\wedge B,\langle w,c,s[y],g[^t/_{|\hat{y}(A\wedge B)|_g^{x,c}s}]\rangle)$
- $Upd(\exists rP,\langle w,c,s,g\rangle)\ =\ Upd(P,\langle w,c,s,g[^r/_{|P|_g^c s}]\rangle)$[81]
- $Upd(\hat{x}A,\langle w,c,s,g\rangle)\ =\ Upd(A,\langle w,c,s[x],g\rangle)$
- $Upd(\iota rP,\langle w,c,s,g\rangle)\ =\ Upd(A,\langle w,c,s,g[^r/_{|P|_g^c s}]\rangle)$

[79] Where z denotes the optional anaphoric context set.

[80] Where I assume that $[[A]]^{w,c,s,g}=1$ iff $[[A]]^{w,c,g}=1$, and where the sequence of variables used for the interpretation of the two conjuncts in a conjunction is the same:

- $[[A\wedge B]]^{w,c,s,g}=1$ iff $[[A]]^{w,c,s,g}=1$ and $[[B]]^{w,c,s,Upd(A,\langle w,c,s,g\rangle)}=1$

We can now represent a sentence with a paycheque pronoun like (57a) by (57b) and (57c), corresponding to sloppy and strict interpretation, respectively, of the paycheque pronoun – that is, as referring respectively to Bill's and John's paycheques:

(57) a. John gave his paycheque to his wife, but Bill gave *it* to his mistress

b. $\hat{x}Gave(x, \iota r\hat{y}PaychequeOf(y,x), x's\ wife)(j) \wedge \hat{x}Gave(x, r(x), x's\ mistress)(b)$

c. $\hat{x}Gave(x, \iota r\hat{y}PaychequeOf(y,x), x's\ wife)(j) \wedge \hat{x}Gave(x, r(j), x's\ mistress)(b)$,

Equally straightforward is the analysis of (58a) and (59a), represented by (58b) and (59b) respectively.[82]

(58) a. Every man lost *a pen*, and some man found *it*.

b. $\forall^v_x[Man(x), \exists y[Pen(y) \wedge lost(x,y)]] \wedge \exists z[man(z) \wedge found(z, y(z))]$

(59) a. Most summers John rents *a car* to go to France. He usually takes *it* on a ferry.

b. $Most^v_x[Summer(x), \exists y[Car(y) \wedge RentIn(j,x,y)]]$
$\wedge\ Usually^x_z[E(z), TakeFerry(j, y(z))]$

We now predict that the only interpretation of (58a), for instance, is that some man who lost his pen found *the unique* pen that he lost.

Before we analyse the other problematic examples discussed above, let us first look at the following sentences:

(60) a. Every soldier deserves a medal. *He* has risked *his* life for *his* country's sake.

b. All soldiers deserve a medal. *They* have risked *their* lives for *their* country's sake.

The pronouns in (60a) and (60b) 'talk about' all soldiers, but certainly in (60a), and optionally in (60b), they are interpreted *distributively*. I will ignore here the non-distributive interpretation of plural pronouns, and propose that plural pronouns can be interpreted distributively by introducing a *distribution operator*, δx, that can front a sentence.[83] Accordingly, I will assume that if A is a sentence, δxA is also a sentence. I will then say that δxA is true in possibility $\langle w, c, g\rangle$ iff for each individual in the set denoted by $g(x)(w)$ it holds that A is true. Thus, δxA is interpreted as follows:

[81] I am assuming here for simplicity that existential quantifiers are still treated as unary and non-anaphoric quantifiers. Nothing hinges on this assumption, however.

[82] From now on I will ignore the distinction between variables and discourse referents, taking a quantified sentence of the form *Every S is P* to be represented by $\forall x[Sx, Px]$ instead of by $\forall x_t[Sx, Px]$ as I would do officially.

[83] See also, among others, van den Berg (1996).

$$[[\delta x A]]^{w,c,g} = 1 \quad \text{iff} \quad \forall d \in g(x)(w) : [[A]]^{w,c,g[^x/_d]} = 1$$

If we now represent the above discourses by the following formula, this discourse receives the desired interpretation:

(61) $\forall^z_x[Soldier(x), \exists y[medal(y) \wedge deserve(x,y)]] \wedge \delta x[Risked-for(x, x's\ life, x's\ cs)]$

Now that we have introduced the distribution operator, we can finally interpret the famous 'telescoping' case of Roberts (1989) and its variant containing plural pronouns:

(62) a. Every player chooses a pawn. *He* puts *it* on square one.

b. $\forall^z_x[Player(x), \exists y[pawn(y) \wedge choose(x,y)]] \wedge \delta x[PutOnS1(x, y(x))]$

(63) a. Every man loves a woman. *They* prove this by giving *them* flowers.

b. $\forall^z_x[Man(x), \exists y[woman(y) \wedge love(x,y)]] \wedge \delta x[PgF(x, y(x))]$

2.10 Donkeys and the specificity of indefinites

In this chapter I have argued that both indefinites and pronouns can be used in two ways.[84] Indefinites can be used specifically and non-specifically, while pronouns can be used referentially and descriptively. A referentially-used singular pronoun can take a specifically-used indefinite as antecedent, and in such a case will refer, if at all, to the unique individual that the speaker of the antecedent indefinite has 'in mind' by his use of the indefinite.

Although the distinction between the two uses of indefinites and pronouns seems very natural, I did make one assumption that might have struck some readers as less plausible. I argued that to account for the universal readings of donkey sentences like the one in (64) we have to assume that the indefinites in the antecedents have to be read specifically, and the pronouns referentially.

(64) If *a farmer* owns *a donkey*, *he* beats *it*.

What I claimed is that we have to look not at the actual, but rather at all *hypothetical* reference contexts to determine the reference of the indefinites and pronouns in donkey sentences like this one. This is because the utterer has specific farmers and donkeys in mind for his use of these sentences. Intuitively, however, this gives rise to a problem. On the one hand, I want to say that an occurrence of an indefinite has a specific speaker's referent in a world w, because the reference context c of the triple $\langle w, c, g\rangle$ assigns to the

[84] In fact I have argued that definite descriptions can also be used in two ways – a natural assumption if pronouns can be used in two ways.

occurrence of the indefinite a specific individual (if any). On the other hand, however, I want to analyse donkey sentences by quantifying over *hypothetical* reference contexts, it must be the case then that world w, together with an assignment function, can also form a triple with many reference contexts distinct from c. But how can such a formalisation capture the intuition that in 'normal' occurrences of indefinites, the indefinite refers to the speaker's referent of the indefinite, the individual that the speaker has in mind?

What is required – as I have shown in section 2.3 – is the assumption that world w can form a triple with many reference contexts (and an assignment function), but that only one of those reference contexts is a *distinguished* one, in the sense that it assigns to indefinites only their speaker's referents in that world.

I believe this is a plausible way to solve the dilemma. It is interesting, though, that with the assumption that pronouns have both descriptive and functional uses, and an appeal to distribution operators like δxA, we can account for donkey sentences without assuming that reference contexts can be shifted and thus that we have to look at hypothetical reference contexts.

Take a look again at our donkey sentence in bishop's clothing (20), repeated here as (65):

(65) If a bishop meets another bishop, *he* blesses *him*.

In section 2.2 I argued that the E-type approach could not account for such sentences, because the pronoun *he*, for instance, obviously could not go proxy for the definite description *the unique bishop that meets another bishop*, because there is no such bishop. But suppose now that we represent (65) by means of the distribution operator 'δx' as follows:

(66) $\exists x[Bishop(x) \wedge \exists y[Bishop(y) \wedge x \neq y \wedge meet(x,y)]] \rightarrow \delta x Bless(x, y(x))$

In this case *he* and *him* could, after all, be treated as descriptive or functional pronouns,[85] and we predict that the sentence means that every bishop who meets another bishop, blesses the unique other bishop whom he meets. Assuming that 'meeting' denotes a non-reflexive relation, (66) does seem to be a natural reading of sentence (65).

This is an encouraging result: if all donkey sentences could be handled in this way, and if in cases of modal subordination it is always descriptive or functional pronouns that count, we can say that uses of indefinites should be represented by eta terms only if (the hearers assume that) the speaker does have a specific individual 'in mind' by his use of the term. This would simplify our analysis considerably, because we would not have to shift the reference context anymore.

Indeed, I believe that this might well be the way to go. However, our analyses still cannot predict the reading that is traditionally assigned to (64): namely that every farmer who owns donkeys beats *every* donkey that he owns. Or better, at this stage I (wrongly?)

[85] This solution is close to the way Neale (1990) and van der Does (1994) account for the universal readings of donkey sentences.

predict that the donkey sentence (64) can be true only if every farmer owns *at most one* donkey.

Fortunately, it is still possible to account for the universal, or unselective, reading of the sentence according to which some farmers own more than one donkey. What we have to do is to generalise the distribution operator in the following way (where $\lambda d.d'$ is the constant function from individuals to d'):[86]

$$[[\delta x, yA]]^{w,c,g} = 1 \quad \text{iff} \quad \forall \langle d, d' \rangle \in [[x]]^{w,c,g} \times ([[y]]^{w,c,g}([[x]]^{w,c,g})) : \ [[A]]^{w,c,g[^{x}/_{d},^{y}/_{\lambda d.d'}]} = 1$$ [87]

Once we have this more general distribution operator available, we can account for the strong, unselective, reading of (64) as follows:

(67) $\exists x[Farmer(x) \wedge \exists y[Donkey(y) \wedge Own(x,y)]] \rightarrow \delta x, yBeat(x, y(x))$

The result will be that for every farmer-donkey pair, it holds that the farmer beats the donkey.

Now a new complication arises. Our distribution operator always has *universal* force. But what happens when we have a donkey sentence with a non-universal adverb of quantification?

(68) Usually, if a farmer owns a donkey, he beats it.

What I will suggest is that adverbial sentences of the form '$ADV_{x,y}(A, B)$', where A contains existential quantifiers associated with variables x and y, will be interpreted as if they were of the form '$A \rightarrow ADV_{x,y}(B)$', where $ADV_{x,y}(B)$ is interpreted as follows:[88]

$$[[ADV_{x,y}(B)]]^{w,c,g} = 1 \quad \text{iff} \quad [ADV](\{\langle d, d' \rangle \in [[x]]^{w,c,g} \times ([[y]]^{w,c,g}([[x]]^{w,c,g})\}, \{\langle d, d' \rangle \in [[x]]^{w,c,g} \times ([[y]]^{w,c,g}([[x]]^{w,c,g}))|\ [[B]]^{w,c,g[^{x}/_{d},^{y}/_{\lambda d.d'}]} = 1\})$$

As a result, a formula like $Usually_{x,y}(\exists x[Fx \wedge \exists y[Dy \wedge Own(x,y)]], Beat(x, y(x)))$ will be true just in case most farmer-donkey pairs that stand in the own relation also stand in the beat relation. At this point, we arrive at a result for (adverbial) quantifiers that we already arrived at for indefinites: namely that we no longer have to shift reference contexts.

Note that if we avoid this shifting of reference contexts, we predict that if an indefinite is used specifically, it always has, so to say, 'wide scope', just as Fodor & Sag (1982) have argued. Their claim is that we should distinguish between referential and quantificational uses of *indefinite* descriptions, on the basis of the fact that the indefinite in a sentence like *John overheard the rumour that a student of mine had been called before the dean* can have not only 'narrow scope' but also maximal 'scope'. The latter possibility is hard to understand if indefinites are treated as quantifiers, because if we replace the indefinite with

[86] See also van Rooy (1998) for a similar rule in a somewhat different framework.

[87] This rule can be extended straightforwardly to an n-ary distribution operator.

[88] This rule can be extended straightforwardly to an n-ary adverbial operator.

a quantificational expression like *each student of mine*, this maximal scope reading is missing. Similarly, the assumption that indefinites can be used referentially seems necessary to account for the fact that we can sometimes refer back in a subsequent sentence to an indefinite occurring in the antecedent of a conditional, as in *If a plumber comes, let him in. He is coming to repair the bathtub.* In general, to account for the 'wide scope' readings of indefinites occurring in so-called 'scope islands', such as modal constructions and *because*- and *if*-clauses, Fodor & Sag propose – just as we did on the basis of anaphora facts – that indefinites can be used not only quantificationally, but also referentially. Unfortunately, according to our analysis of referentially- or specifically-used indefinites, so-called *intermediate* readings are predicted to be impossible. However, as has been shown by Abusch (1993), among others, this prediction is wrong: indefinites can also escape from islands to yield intermediate readings in constructions like the following ones:

(69) a. Every professor rewarded every student who read *a book he had recommended.*

b. Each choreographer believes that it would be damaging for *a dancer of his* to quit the company.

If we wanted, we could account for wide and intermediate readings of indefinites by giving the terms wide or intermediate scope, and stipulate that terms can behave differently from quantifiers under islands. But perhaps this scope analysis is too stipulative since it offers no reason why terms can behave differently from quantifiers under islands. And anyway, once we assume that reference contexts don't shift, anymore, the relative scope of the indefinite term becomes (almost) irrelevant for its interpretation. On the one hand, this is a very nice feature of our new analysis, since it gives us an *explanatory* connection between the properties of indefinites and both their unusual behaviour with respect to *scope* and their ability to figure as the antecedent of singular *anaphoric pronouns* in subsequent sentences. But then the problem remains of how to account for the *intermediate* readings mentioned above. We can account for intermediate readings within the present framework by taking over Kratzer's (1998) proposal to make use of Skolem functions. The idea is that speakers can have in mind in their use of indefinites not only specific individuals, but also specific Skolem functions, functions from an n-tuple of individuals to individuals. The intermediate reading of (69a), for instance, can then be accounted for by assuming that the speaker, in using the indefinite *a book he had recommended*, has the particular Skolem function in mind that assigns to every (relevant) professor a particular book that he had recommended.

Notice that an interesting result emerges when we no longer assume a shifting of reference contexts when we interpret negated and quantified sentences and conditionals; and when non-specifically used indefinites introduce not specific objects but also properties into the discourse, and specifically used dependent indefinites introduce only Skolem functions. This is that all of the operators that in DRT/FCS figure as plugs with respect to anaphoric binding can now simply be treated as holes. We might redefine the change function of negation, for instance, as follows:

- $Upd(\neg A, \langle w, c, g\rangle) \quad = \quad Upd(A, \langle w, c, g\rangle)$

In this book I will make no further use of Skolem functions. However, the assumption that operators that are absolute plugs with respect to anaphoric binding in standard DRT/FCS/DPL are, in fact, 'leaky' – negation being only one example of such an operator – will play an important role in the analysis to be offered in the next chapter.

Chapter 3

Intentional Identity

3.1 Introduction

According to the received view in semantics, so-called unbound pronouns – pronouns not bound by a quantifier Q inside the smallest clause containing Q – should be treated either as abbreviations for the antecedent clause or as variables bound by a (dynamic) existential quantifier. Geach's notorious Hob-Nob sentences, exemplifying intentional identity attributions, have always been a threat to this assumption.

In this chapter, I relate the problem that Geach's Hob-Nob sentences pose for the traditional analyses of pronouns to the problem that examples of *pronominal contradiction* pose for the same theories. My proposal for solving Geach's problem will be similar to my proposal for how to account for pronominal contradiction as presented in the previous chapter, and likewise will involve taking the notion of *speaker's reference* seriously in dynamic semantics.

This chapter will serve not only to give an additional argument for taking this notion more seriously than is usually done, but also to discuss some other issues. First, I will point out both the *similarities* and the *differences* between *intentional identity* attributions (or *Hob-Nob sentences* as I will call them), on the one hand, and cases of *information exchange* or *Hob-Nob situations*), on the other.[1] In doing so, I will also address the question of how far intentional identity attributions suggest that belief states should be structured around *belief objects*.

In section 2 of this chapter, I will discuss Geach's traditional problem of intentional identity and Edelberg's (1986) more recent asymmetry problem. In section 3, I will consider how to account for Edelberg's asymmetry problem following standard approaches towards anaphoric dependence, by relating intentional identity attributions to cases of information exchange. In section 4, however, I show why these obvious proposals won't work in general. The problems can be solved, however, if we take the notion of speaker's reference seriously, as I discuss in sections 5 and 6. In the last section I will briefly address the above mentioned

[1] See also Dekker & van Rooy (1998).

ontological question of whether or not intentional identity attributions suggest that belief states should be structured around belief objects.

3.2 The problem of intentional identity

A key problem that every semantic account of anaphora faces is that a pronoun occurring in the embedded clause of an attitude attribution can have as its syntactic antecedent an indefinite in the embedded clause of an earlier attitude attribution. In a logical language this is not difficult to represent if the indefinite is interpreted *de re*. But the problem is that this doesn't always seem to be the case. This is the problem discussed under the heading of *intentional identity* by Geach (1967), and called the problem of *de dicto pronouns* by those working in the tradition of Montague semantics. Examples of these sentences include the following:

(70) John believes that *a woman* broke into his apartment.
He believes that *she* is now hiding from the police.

(71) Carl wants to catch *a fish* today, and he wants to eat *it* afterwards.

(72) Hob believes that *a witch* blighted Bob's mare,
and Nob believes *she* killed Cob's sow.

On the intended readings of these sentences, the attitude attributions can be true without there being a woman about which John has beliefs, a fish that Carl wants, or a witch that is responsible for the beliefs of Hob and Nob. For (72) to be true, there does not even have to be an existing individual that is the focus of both Hob's and Nob's beliefs. This is shown by the following Geachian story:

> Last night, Bob's mare became quite ill. Hob, who tends Bob's barn, inferred that a witch blighted her. This morning Hob said to his friend, Nob, "A witch blighted Bob's mare." Nob believes what Hob has told him. He thinks for a moment, and says, "Cob's sow died early this morning. I'll bet the same witch killed the sow, too." But in fact both animals fell ill due to perfectly natural causes. (Edelberg, 1986, pp. 1-2)

According to this story, the Hob-Nob sentence (72) would be true. In the Geachian tradition, anaphoric elements are treated as bound variables; but the problem is that there is no way to bind the variable that represents the pronoun in the second clause by the quantifier that represents the indefinite in the first clause if you can quantify only over existing individuals. In the framework of traditional Montague semantics, the following translations might be tried (where h stands for 'Hob', n for 'Nob', BBM for 'Blighted Bob's mare', and KCS for 'killed Cob's sow'):

(73) a. $Bel(h, \exists x[witch(x) \wedge BBM(x)]) \wedge Bel(n, KCS(x))$

b. $Bel(h, \exists x[witch(x) \wedge BBM(x) \wedge Bel(n, KCS(x))])$

c. $\exists x[witch(x) \wedge Bel(h, BBM(x)) \wedge Bel(n, KCS(x))]$

If pronouns are treated as bound variables, it seems that the only possible way to go is to use either representation (73b) or (73c). Unfortunately, (73b) doesn't give the intended reading because the attitude attribution doesn't seem to say anything about what Hob believes about Nob's beliefs, and representation (73c) does not predict (72) to be true in the above story because, in fact, witches do not exist.

From these problems some have concluded that variable x should really range over *non-existent* objects, and that cases of intentional identity should be translated as in (73c) after all. In cases of intentional identity, a *de re* belief attribution is made about a specific object that might be non-existent. One problem with this assumption is that a sentence like (71) doesn't seem to be about a specific (maybe non-existent) fish at all. There does not need to be one specific fish that Carl's belief is about such that Carl believes he will catch it and wants to eat it afterwards to make the attitude attribution true.[2] Let's call this problem the *specificity problem*. In addition, for (72), for instance, to be true, it should be predicted that in all of Hob's belief alternatives there is a witch who blighted Bob's mare, something that is not guaranteed if we represent (72) by (73c). These two problems suggest that we should represent intentional identity attributions in a non-Montagovian way, as in (74):

(74) $\exists x Bel(h, W(x) \wedge BBM(x)) \wedge Bel(n, KCS(x))$

In fact, Slater's proposal (1988) boils down to this. According to this proposal, Hob and Nob have a belief about a specific object, but all we know about this object is that Hob thinks that it is a witch that blighted Bob's mare and Nob believes that it killed Cob's sow. But intuitively (72) can be true without any specific object satisfying the above conditions. The reason is that there need not be one actually-existing object that is responsible for the relevant beliefs of Hob and Nob. Hob believes of none of the individuals he has ever come across to be a witch; thus none of them satisfies the property expressed by $\hat{x}Bel(h, W(x) \wedge BBM(x))$ (cf. Buridan, 1350). Arguing that variables should also range over non-existing objects does not help if it is assumed that indefinites occurring in the embedded clauses of belief attributions will be represented by a formula where the corresponding existential quantifier has wide scope with respect to the belief predicate. This would give rise to the unwanted prediction that for the first conjunct of (71) to be true, there must be a specific object about which Carl has the belief that it is a fish that he will catch today.

[2]See also Haas-Spohn (1986)

All of these problems suggest that we should indeed represent a sentence like (72) by (74); but that the variables should range not over specific objects, but over *individual concepts* instead. Something like this was proposed by Saarinen (1978) to account for intentional identity attributions.[3] He assumed that variables range over individual concepts and that these concepts don't have to be instantiated in the actual world. However, as shown by Edelberg, this suggestion is problematic. If we don't restrict the range of the variables, Saarinen's proposal would predict that attributions of the form (75) are equivalent to attributions of the form (76):

(75) $\exists x Bel(a, Px) \wedge Bel(b, Qx)$

(76) $\exists x Bel(b, Qx) \wedge Bel(a, Px)$

However, Edelberg (1986, 1992, 1995) observed that intentional identity attributions are in general *not symmetric*. Consider the following case:

> Arsky and Barsky investigate the apparent murder of Smith, and they conclude that Smith was murdered by a single person, though they have no one in mind as a suspect. A few days later, they investigate the apparent murder of a second person, Jones, and again they conclude that Jones was murdered by a single person. At this point, however, a disagreement between the two detectives arises. Arsky thinks that the two murderers are completely unrelated, and that the person who murdered Smith, but not the one who murdered Jones, is still in Chicago. Barsky, however, thinks that one and the same person murdered both Smith and Jones. However, neither Smith nor Jones was really murdered. (Edelberg, 1995, p. 317)

For this case we find (77) but not (78) acceptable:

(77) Arsky believes that someone murdered Smith, and
Barsky believes he murdered Jones.

(78) Barsky believes that someone murdered Jones, and
Arsky believes he murdered Smith.

Intentional identity attributions are in general *not symmetric*, although Saarinen's proposal wrongly predicts them to be. Edelberg called this problem the *asymmetry problem about intentional identity*. Note, too, that any proposal that seeks to account for intentional identity by representing sentences like (72) by (74) and by allowing quantification over non-existing objects fails to explain this asymmetry.

A different but related problem is discussed by Edelberg under the heading of the *variable aboutness problem of attitudes de re*. The problem is related to the following case:

[3] See also Zeevat (1996).

> Smith and Jones are dead. A single person murdered both of them. Detective Arsky investigates both cases, and comes to believe that someone murdered Smith and that someone murdered Jones, but he doesn't have anyone in particular in mind as a suspect. Arsky does not believe that Smith's murderer and Jones's murderer are the same person. (Edelberg, 1995, p. 318)

The problem is to account for the intuition that on their most straightforward readings, (79) is true, while (80) is false:

(79) Someone murdered Smith, and Arsky thinks he didn't murder Jones.[4]

(80) Someone murdered Smith, and Arsky thinks he murdered Jones.

The problem for an approach on which variables range over concepts is that such an approach predicts that (80) as well as (79) is true, because there is a single concept, *the murderer of Jones*, whose instantiation in the actual world murdered Smith and whose instantiation in Arsky's belief worlds also murdered Jones in each of them.

Now we have three kinds of problems. First, we have cases like (70) and (71), where only one agent is involved and the pronoun in the second sentence does not refer back to a specific existing object that the speaker refers to. Second, we have *de re* attributions like (79) and (80), where the pronoun in the second sentence *does* refer back to such a specific existing object. And third, we have intentional identity attributions like (72), where two agents are involved and the pronoun does not refer, for the speaker, to a specific existing individual. For *de re* attributions we have to account for the variable aboutness problem; and for intentional identity attributions with more agents involved, we have to account for the asymmetry problem.

In the previous chapter we discussed several frameworks that can handle anaphoric dependencies across sentential boundaries. It is only to be expected that the intentional identity cases discussed above could be handled in one of these frameworks, too. In fact, this is what I believe. But as we will see, the solution to these intentional identity problems is not as straightforward as one might hope.

3.3 Asymmetry explained by descriptive approaches

Standard dynamic semantics, or CCT, as discussed in the previous chapter, has become a very popular way to account for anaphoric relations across sentential boundaries. According to this theory, each sentential clause/formula is interpreted with respect to a unique context, where this context represents information about the subject matter of conversation and the values of variables. Whereas in traditional semantic theories the primary goal was to determine the *truth conditions* of sentences in a systematic way, in these more

[4] I am making use of this rather awkward phrasing to keep scope matters clear.

recent theories more attention is paid to the ways in which sentences *change the context* of interpretation.

In CCT, contexts are typically represented by sets of world-assignment pairs. In this way, a context can represent not only the 'world' information about what is presupposed with respect to the subject matter of conversation, but also the information about the possible values of variables, or *discourse referents*. If we fix a world, and concentrate only on the latter kind of information, we can represent a context by a set of (partial) assignment functions. An indexed sentence like

(81) A man_x is walking in the park

will now update a context, C, by *enriching* the assignments of this context; each new assignment will also assign a value to a variable (or discourse referent) x, and *each* man who is walking in the park in this fixed world will be the value of variable x with respect to one of the assignments of the updated context. Thus, the only information associated with x in this new context, C', is that the value of x is a man who is walking in the park in this world. As a result, a subsequent sentence like

(82) He_x is whistling

can now be interpreted with respect to this updated context. If we again fix a world, sentence A, according to the above theories, will be true in this world with respect to assignment g iff the update of $\{g\}$ with A, $[[A]](\{g\})$, is non-empty. Similarly, the discourse $A_1, ..., A_n$ will be true with respect to assignment g iff $[[A_n]](...([[A_1]](\{g\}))...)$ is non-empty. As a result, the discourse (81) - (82) is predicted to be true iff there is a man who is walking in the park and is whistling.

Note that these theories predict that in our above sequence the pronoun *he* is an abbreviation for the *indefinite* description *a man who is walking in the park*, which is recoverable from the antecedent clause. The reason is that the only information associated with variable or discourse referent x in the context resulting from the update of the first sentence is that the value of x is *a* man who is walking in the park.

3.3.1 Cross-speaker anaphora

Although CCT has been developed to account for anaphoric and presuppositional dependencies in discourses made by a single speaker, it seems we can also use CCT to analyze cases in which two or more agents exchange information about an object. I will call such situations of information exchange *Hob-Nob situations*. Typically these involve the use of pronouns by one agent to refer back to objects mentioned or introduced by another agent. Hence the term "cross-speaker anaphora", which is also used. Consider the following dialogue between Arsky and Barsky:

(83) Arsky: *Someone* murdered Smith.
Barsky: *He* also murdered Jones.

Standard dynamic semantics seems able to account reasonably well for what is going on in these sentences. First, it seems obvious that for both sentences to be true, it has to be the case that someone who murdered Smith also murdered Jones, just as the above definition of the truth conditions of discourses predicts. Thus, if we want to be able to determine the truth conditions of the second sentence relatively independently of what is asserted by Arsky we can treat the pronoun *he* as an abbreviation for the *indefinite* description *Someone who murdered Smith*.

Second, it seems reasonable to make the Gricean assumption that if somebody makes an assertion, he should also *believe* what he asserts. But in normal cases there seems to be an *asymmetry* between what Arsky has to believe and what Barsky has to believe in order to make their respective assertions appropriately. In normal cases we infer that Arsky has only to believe the content of what he asserts himself: that there is someone who murdered Smith; while Barsky can use the pronoun appropriately only if he also believes what is asserted by Arsky. That is, Barsky has to believe that there is someone who murdered both Smith *and* Jones. This asymmetry can be readily explained on standard dynamic semantics (see Groenendijk et al., 1997) if we make one extra assumption. This is that if a speaker does not respond, we can assume that he has accepted, and thus believes, what has been asserted by the earlier speaker. With the help of this assumption we can infer for this situation of information exchange, i.e. a Hob-Nob situation, that Barsky believes, after the update of his belief state with what is asserted by Arsky, that there is someone who murdered Smith.

Note that a similar explanation can be given for the asymmetry in (83) if we assume that the pronoun is an E-type pronoun and is used as an abbreviation for the *definite* description recoverable from the antecedent sentence. In this case we would predict that for Barsky to make his assertion appropriately, he has to believe that *the* one who murdered Smith also murdered Jones.

In the ideal case, both what is expressed by the second speaker and the asymmetry between what it is necessary for Arsky and Barsky to believe for each of them to make their assertions appropriately in the above discourse, can be explained straightforwardly by means of both standard dynamic semantics and the E-type approach. But standard dynamic semantics is a bit more general; it can also explain similar cases of asymmetry where no pronouns are involved. Given the very similar behaviour of pronouns and *presupposition triggers*,[5] we can expect the same pattern for presupposition triggers. And indeed that is what we find in cases like the following one (where $[\text{My}]_F$ indicates that *my* has focal accent):

[5]cf. Kripke (ms) and van der Sandt (1992).

(84) John: My parents are gone.
Mary: [My]$_F$ parents are gone *too*.

First, when we interpret the second sentence with respect to a context updated by what John said, this context will *satisfy* the presupposition triggered by *Mary*'s utterance that the parents of somebody other than Mary are gone. Second, from Mary's utterance (but not from John's), we infer that Mary (but not John) believes that both her and someone else's parents are gone.

3.3.2 Intentional identity

Notice that the asymmetry between what the first and the second speaker typically have to believe to make their respective assertions appropriately in the above cases of *information exchange* is very similar to the asymmetry between what is ascribed to Arsky and to Barsky, respectively, in the *intentional identity* attributions (77) and (78) discussed in section 2. This suggests that we can also account for the asymmetry that shows up in intentional identity attributions, or Hob-Nob sentences as I will call them, with either an E-type approach or a straightforward extension of standard dynamic semantics. And anyway, in intentional identity attributions we have to deal with anaphoric dependencies across sentence boundaries – a phenomenon for which the E-type approach and dynamic semantics were invented – if these sentences are to be interpreted in an *incremental* way.

Given my analysis of descriptive pronouns in the previous chapter, it should be clear how the E-type approach would account for both (i) the possibility of anaphoric dependencies across belief attributions, and (ii) the observed asymmetry. Notice that the E-type approach can also account straightforwardly for the variable aboutness problem of attitudes *de re*, as exemplified by the truth of (79) and falsity of (80).

It is also easy to imagine how CCT should be extended to account for the possibility of intentional identity attributions and their observed asymmetric behaviour. We have seen above that on a dynamic semantic account every sentence (i) is interpreted with respect to a context represented by a set of world-assignment pairs; and (ii) *creates* a new context, the context resulting from the earlier context updated by the current sentence. Later sentences can then be interpreted with respect to this later context. The idea now is to do something similar for *embedded* clauses in attitude ascriptions in the case of intentional identity attributions. The only difference is that embedded clauses in attitude attributions do not have to be interpreted with respect to the *main* context, but only with respect to a *subordinated* context; and they create contexts with respect to which only subsequent embedded clauses, rather than entire assertions, have to be interpreted. Accordingly, embedded clauses should not be interpreted with respect to the main context, but rather with respect to subordinated contexts introduced into the discourse by the interpretation of an earlier embedded clause.

As it happens, Geurts (1995, 1998) has already made use of this modal subordination

approach to account for anaphoric and presuppositional dependencies in attitude attributions for the one-agent case. In his analysis, formulae representing attitude attributions are interpreted with respect to *old* information states and set up *new* ones. These old and new information states are then indexed by *propositional discourse referents*. Thus the intentional identity attribution like (70), repeated here as (85), can be represented by the formula in (86):[6]

(85) John believes that *a woman* broke into his apartment.
He believes that *she* is hiding from the police.

(86) $Bel^q_r(j, \exists x Wx \wedge B\text{-}in\text{-}Ax) \wedge Bel^r_s(j, HPx)$,

Here q denotes the context of interpretation with respect to which the embedded clause $\exists x Wx \wedge B\text{-}in\text{-}Ax$ is interpreted, and r the newly introduced subordinated context. The context denoted by r will contain information about the variable x, and associates with it the information that it is a woman who broke into John's apartment. Because the second embedded clause is interpreted with respect to this newly created context, the pronoun, represented by a free variable, can be interpreted. To determine whether or not a belief attribution represented by $Bel^q_r(j, A)$ is true in a given world or not, Geurts assumes that the belief state of an agent is represented by a set of possible worlds, and that the above formula is true in w iff for every world v consistent with what a believes in w, there is an assignment h such that $\langle v, h\rangle$ is an element of the context denoted by q updated by A.

Geurts uses his framework to account only for *single* agent cases of intentional identity attributions, but of course we might use his analysis for *multi*-agent cases too. Note that if we do so, we can immediately explain the asymmetry between (77) and (78), represented here schematically by (87) and (88) respectively:

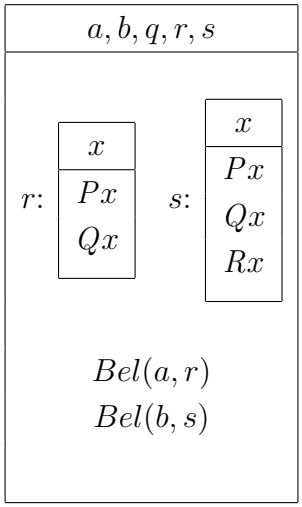
(87) $Bel^q_r(a, \exists x Px \wedge Qx) \wedge Bel^r_s(b, Rx)$

(88) $Bel^q_r(a, \exists x Rx) \wedge Bel^r_s(b, Px \wedge Qx)$

If we switch to discourse representation structures, we can say what the logical form (87) amounts to in these terms:

[6] I use the FCS/DPL framework rather than the DRT framework that Geurts uses. Although the choice of framework (representational or not) is important for Geurts' analysis of presuppositions, it is not crucial for the examples that we will discuss.

(89)



In this way we predict that (i) the pronoun *he* in (77) is an abbreviation for the indefinite description *someone who murdered Smith*, and that (ii) (77) is true in a situation where Barsky, but not Arsky, has a one-murderer theory, i.e. believes that the same person murdered both Smith and Jones.

Geurts (1995) assumed that we should introduce propositional discourse markers only when *embedded* clauses are interpreted. But, as we have seen in chapter 2, if we also treat presupposition as a propositional attitude, we might introduce a (distinguished) propositional discourse marker, p, that represents what is presupposed in each possibility. What is important is that once we assume that possibilities also contain the information that is presupposed, we can also account straightforwardly for the variable aboutness problem of attitudes *de re* as discussed in section 2. Remember that p denotes what is presupposed in the main context. If we represent (79) and (80) by (90) and (91) respectively, we predict correctly that (79) is true and (80) false, in the situation described in section 2 of this chapter, because Arsky does not have a one-murderer theory.

(90) $\exists x MSx \wedge Bel^p_q(a, \neg MJx)$

(91) $\exists x MSx \wedge Bel^p_q(a, MJx)$,

Just as in the above Hob-Nob situation, also for Hob-Nob sentences the dynamic semantic solution is more general than the E-type approach. By extending dynamic semantics as above we can explain not only this asymmetry with respect to pronouns, but also when (other) *presupposition triggers* are involved. Consider the following example, adapted from Heim (1992):

(92) a. John is sure that his parents are gone.

b. Mary thinks that [her]$_F$ parents are gone, too.

In an utterance of (92b), with focus accent on *her*, it seems that *too* may relate to the information that John's parents are gone and not to the information that John thinks that

his parents are gone. On such an analysis, I think, it need not be presupposed that John's parents are gone, but the sentence gives rise to the expectation that Mary believes that John's parents are gone. Notice that this expectation can be explained straightforwardly by means of modal subordination.

3.4 Problems for descriptive approaches

In the previous section we saw how both the E-type analysis of pronouns and (a straightforward extension of) standard dynamic semantics can account for the asymmetry between what Arsky and Barsky have to believe (i) in order for them to make appropriate assertions when they are engaged in a conversation, and (ii) to account for the fact that the belief attribution (77) is true on its most straightforward reading, while (78) is false. Indeed, it seems that the two theories make pretty good predictions. But the problem is that this is the case only if certain *ideal conditions* hold. Unfortunately, ideal or normal conditions do not always obtain.

3.4.1 Cross-speaker anaphora

Consider first the case of information exchange, or Hob-Nob situations. Ideal conditions need not obtain here, for instance, because the following dialogue involves a perfect exchange of information, even if there is no man running through the park:

(93) A: *A man* is running through the park.
B: *He* wears Nike sport-shoes.

Of course, what A has to believe to make his assertion appropriately still has to be the same as in the ideal case; and of course, in this non-ideal situation, the first sentence, and thus the whole discourse, will not be true.

What is interesting, though, is that whether or not B believes or accepts what A says, it seems that what B asserts himself can be true, even if what A says is false. If this is indeed the case, we can conclude that personal pronouns cannot be treated simply as abbreviations for (in)definite descriptions recoverable from the antecedent indefinite.

To make these cases clearer, let's look at an example of *pronominal contradiction*, already discussed in chapter 2. Consider the following dialogue:

(94) A: A man is running through the park.
B: *He* is not a man, but just a boy, and *he* is not running, but just walking.

Such examples differ from the ideal case in two ways: First, although B is saying something coherent, we cannot determine the proposition expressed by him by treating the pronouns as abbreviations of the description *a/the man who is running through the park*, for that would give rise to the impossible proposition. Second, to be able to make this assertion

appropriately, B also cannot *believe* that the 'referent' of the pronoun is a/the man who is running through the park.

So, although the *truth value* and the *appropriateness* of what B asserts are dependent somehow on A's speech act, this dependence cannot be explained in the most obvious way known from the E-type approach or from dynamic semantics. That is, we cannot interpret B's assertion as a *monotone* update of an initial context updated by A's assertion.

At first sight it might seem obvious how to handle cases of pronominal contradiction in standard dynamic semantics. According to these theories, we do two things when we update an initial context with what is asserted by A: (i) we introduce a discourse referent induced by the indefinite *a man*, and (ii) we associate with this discourse referent the descriptive material "being a man walking in the park". In terms of Discourse Representation Theory (DRT), this would result in the following DR-structure:

(95)
x
$Man(x)$
$Running-through-park(x)$

When a second speaker uses a pronoun whose denotation depends on the indefinite used by the first speaker, but denies the descriptive material associated with it, we might say that B's assertion that it is a boy who is walking in the park should be interpreted with respect to the earlier context from which the descriptive material has been eliminated.

(96)
x
$\vdots$

Of course, no proponent of standard dynamic semantics has ever made this proposal. The reason is obvious: the only information that these theories associate with a discourse referent is the *existential* information that something exists. But this information will not be enough to explain the *appropriateness* of the dialogue in (94).

3.4.2 Intentional identity

Ideal conditions do not always obtain in intentional identity attributions either, as observed by Geach (1967). Geach only discussed an analysis of pronouns as abbreviations for *definite* descriptions recoverable from the antecedent clause, but his argument immediately carries over to its *indefinite* counterpart. Geach argued against the descriptive approach because the second agent need not believe all of the descriptive material recoverable from the antecedent sentence. In Geach's original sentence, for example, Nob doesn't have to believe

that the witch that he is thinking about blighted Bob's mare, nor that Hob believes this. It seems that intentional identity attributions can be truly and appropriately made even if the agents disagree about the descriptive content associated with the belief attribution.

> The Gotham city newspapers have reported that a witch, referred to as "Samantha", has been on quite a rampage. According to the article she has been blighting farm animals and crops and throwing people down wells. In reality, there is no such person: the animals and crops all died of natural causes, and the people found at the well-bottoms had all stumbled in by accident in a drunken stupor. The news reporters simply assumed that a witch was responsible for all the mishaps, and dubbed her "Samantha". Hob and Nob both read the Gotham Star and, like most folks, they believe the stories about the witch. Hob thinks Samantha must have blighted Bob's mare, which took ill yesterday. Nob thinks Samantha killed his friend Cob's sow. Nob has no beliefs at all about Hob or about Bob's mare; he is unaware of the existence of either. (Edelberg, 1986, p. 2)

Note how similar this problem of intentional identity attributions is to the problem of pronominal contradiction which plagues classical dynamic semantics. In neither case can pronouns be treated as in the popular analyses of pronouns, as abbreviations for the indefinite or definite descriptions recoverable from the antecedent clause. And just as in the pronominal contradiction case, the most obvious move here is to interpret the second embedded clause not with respect to the context resulting from the update of the *whole* of the first embedded clause, but with respect to the context resulting from the update of only *part* of the first embedded sentence. Instead of representing the intentional identity attribution (72) by something like

(97) $Bel^q_r(h, \exists x Px \wedge Qx) \wedge Bel^r_s(n, Rx)$,

we can now represent it by

(98) $Bel^q_r(h, \exists x Px) \wedge Bel^r_s(h, Qx) \wedge Bel^r_t(n, Rx)$.

As a result, we predict that Nob does not have to believe everything that is attributed to Hob, just as we want.

But, as in the case of pronominal contradiction, the move won't work here either. The reason is the same in both cases: according to standard dynamic semantics the only information associated with a discourse referent is *existential* information, which is *too weak* to account for the data. In this case it is too weak because it makes belief attributions too easily true. Because Nob agrees with almost nothing that is attributed to Hob, almost none of the descriptive material occurring in the embedded clause of what is attributed to Hob can occur in the description that the pronoun is going proxy for. As a result, the

indefinite description will not be much richer than *someone*, which can hardly be enough to explain *why* the intentional identity attribution could be used appropriately in the first place.[7]

3.5 Speaker's reference

As discussed in chapter 2, a more natural way to account for the phenomenon of pronominal contradiction is to assume that pronouns can at least sometimes be used *referentially*, referring back to the individual that the first speaker had in mind for his use of the antecedent indefinite. We called this latter individual the *speaker's referent* of the (use of) the indefinite.

At this point the obvious claim to make would be that the notion of speaker's reference is also crucial to account for the appropriate use of many Hob-Nob sentences. Indeed, this is what I want to propose.[8]

It should be clear, however, that to account for intentional identity attributions, the notion of speaker's reference cannot be cashed out in exactly the same way as we did before. In the previous chapter I assumed that an occurrence of a specifically-used indefinite introduces with respect to each reference context a specific real existing individual to the discourse. This assumption, however, must be given up. But we can generalize our analysis and say that indefinites introduce not specific *individuals*, but rather specific *individual concepts* into the discourse. Thus, the reference-contexts used in section 3 of chapter 2 should no longer be functions from indices to individuals, but rather functions from indices to individual concepts. Similarly, assignment functions will no longer be functions from variables and discourse referents to individuals, but instead functions from variables to individuals and from discourse referents to individual concepts. This has no major consequences for the formal analysis; we only have to slightly re-define the interpretation rule for terms.[9]

One way to go, now that we have changed the formal objects introduced by specifically-used indefinites, would be to say that the indefinite antecedents used in Hob-Nob sentences will have *wide scope* with respect to the verb of belief. But as noted already in the second section of this chapter, this solution seems problematic; this is because for a sentence like (72) it does not guarantee that in all of Hob's belief alternatives there is a witch who blighted Bob's mare. We saw in section 2 that Saarinen (1978) proposed to solve this

[7]The intentional identity example for the straightforward modal subordination account also has its presuppositional counterpart (see Dekker & van Rooy (1998)).

[8]See also van Rooy (2000).

[9]The interpretation rule for terms should be re-defined as follows (ignoring iota-terms):

- $[[t]]^{w,c,g} = g(t)(w)$, if t is a variable,
  $= d$, if $t = \eta r_n P$, $c(n)(w) = d$ and $d \in I_{w,c,g}(P)$,
  $= *$ otherwise

problem, on the assumption that indefinites should be treated as existential quantifiers, by splitting the existential quantification itself from the descriptive contribution (that of being a witch).

I will not assume, however, that the 'contribution' of indefinites to the discourse should be split in this way. I will give the indefinites used for intentional identity attributions *narrow scope* with respect to the belief predicate. But because I will assume that when we interpret a belief attribution in possibility $\langle w, c, g\rangle$ we should analyze the indefinites occurring in the embedded clause with respect to reference context c, what I propose will turn out to be roughly equivalent to the '*wide scope + split*' analysis suggested above.

If we take the notion of speaker's reference seriously, and also don't assume that the belief relation is a plug with respect to anaphoric binding, we no longer need modal subordination to account for intentional identity attributions. That is, the contexts that are introduced and picked up by belief attributions should simply be represented by possible worlds. When we now interpret a clause of the form $Bel^q_s(a, A)$ as given below, it is easy to see that the individual concept introduced by the eta term $\eta r_n P$ will depend on the *actual* reference context, and that this concept can be picked up by a pronoun in a later belief attribution.

- $[[Bel^q_s(t, A)]]^{w,c,g} = 1 \quad \text{iff} \quad \forall v \in K([[t]]^{w,c,g}, w) :\ v \in g(q)\ \&\ [[A]]^{v,c,g} = 1$
- $Upd(Bel^q_s(t, A), \langle w, c, g\rangle) \quad = \quad Upd(A, \langle w, c, Upd(t, \langle w, c, g\rangle)\rangle)[^s/_{\{[[A]]^{v,c,g} | v \in g(q)\}}]$

Now we can represent our problematic intentional identity ascriptions as follows:

(99) $Bel^q_s(a, Q(\eta r_n P)).\ Bel^l_m(b, Rr)$,

In this representation the subordinated contexts doesn't have to play a role (if $s \neq l$): the anaphor can take the indefinite of the first ascription as antecedent, though it need not be the case that the referent has property P. As a result, our problematic Hob-Nob sentence is predicted to be true when (i) Hob has the *existential* belief that there is a witch that blighted Bob's mare, and (ii) that the instantiation of the *specific* concept that the *speaker* had in mind for his use of the antecedent indefinite killed Cob's sow in each of Nob's belief alternatives.

Notice that the problem that arose for the earlier analysis does not arise now. The earlier analysis didn't work because it was *too weak*, making belief attributions too easily true. Because Nob agrees with almost nothing attributed to Hob, almost none of the descriptive material occurring in the embedded clause attributed to Hob can occur in the description that the pronoun is going proxy for. As a result, the indefinite description will not be much richer than *someone*, which doesn't explain the appropriate use of the intentional identity attribution. When we assume that speaker's reference is relevant here, we can explain why the intentional identity attribution can be used appropriately in such cases after all. The reason is that we can now associate more than just existential information with a discourse referent; the information associated with the discourse referent for

our Hob-Nob sentence (72) can now be something like 'the individual of world v that plays in that world the "Samantha"-role'.

Crucial for our analysis is that the individual concept introduced by specifically-used indefinites depends on the *actual* reference context. At first sight, this might just seem like a trick. However, the proposal is really based on a very natural assumption. When we make use of modal subordination to interpret embedded sentences in a discourse, we assume that different kinds of contexts, or information states, 'exist' in the discourse, and can be exploited to interpret sentences. One of these contexts, the *main* context, is the one that represents what is presupposed by the speaker and the other participants in a conversation. The other contexts are *derived* or *subordinated* contexts; these contexts are used to account, among other things, for the anaphoric and presuppositional dependencies between the relevant embedded clause, on the one hand, and (embedded) sentences used earlier in a discourse on the other.

It seems that we should represent all subordinated contexts in the same way as we do the main context. Indeed, this was what the modal subordination proposal discussed in section 3 amounted to. Where the main context represents the information that the *participants in the conversation* believe and presuppose, subordinated contexts used for the analysis of belief attributions should represent what the relevant *agents* believe and presuppose. And isn't this exactly what the similarity between Hob-Nob situations and Hob-Nob sentences suggests? Note, however, that while I have stressed the similarity in the previous sections between Hob-Nob situations and Hob-Nob sentences, there is also an important *difference*. The difference is that if Arsky and Barsky are engaged in a conversation, it is Arsky and Barsky *themselves* who are *responsible* for their use of pronouns and presupposition triggers; but that when a speaker *attributes* beliefs to Arsky and Barsky or to Hob and Nob, it is *the speaker* who is *responsible* for the anaphoric and presuppositional links, and not the agents that the belief attribution is about.[10] This difference in responsibility for the relevant link can be modelled by a difference between the kinds of information that main and subordinated contexts might contain.

Some subordinated contexts will be *consistent* with the main context, and might even contain *more* information. For instance, the context of interpretation of the consequent of an indicative conditional will be the subordinated context resulting from adding the information of the antecedent to that of the main context. Other subordinated contexts might not simply be extensions of the main context, but might *lack* certain *information* that the main context contains. The subordinated contexts used for the analysis of belief attributions typically belong to the latter category.

One of the essential pieces of information that the main, or basic, context contains is that the speaker and the other participants in the conversation always inhabit all of the worlds/possibilities of this context; the speaker and the other participants do not know which world they are in, but at least presuppose that they exist and that the actual dis-

[10]See Dekker & van Rooy (1998).

course is taking place. A subordinated context used for the interpretation of the embedded clauses of attitude attributions need not contain this information, for we can attribute a belief to an agent who doesn't know of our existence, or that we are making this attitude attribution.[11] As a result, in analysing belief attributions, for instance, we should not treat a subordinated context as a context in which the agent himself is uttering the embedded clauses that the speaker uses in the belief attribution.

Normally, Hob-Nob sentences are used to describe Hob-Nob situations, and in these cases the difference between main and subordinated context is not crucial. However, Hob-Nob sentences can also be used to describe situations that are not Hob-Nob situations. Such cases are typically problematic for descriptive approaches to intentional identity attributions, as shown in the previous section. To implement the fact that the *speaker* and not the relevant agent(s) is responsible for the anaphoric dependencies in intentional identity attributions, I have assumed that it is the *actual* reference context that determines which individual concept is introduced into the discourse, and not the reference contexts that are consistent with what the agents themselves believe.

With this intuitive underpinning for our formal analysis, it seems reasonable to assume that when c is the actual reference context, v is a world consistent with what the relevant agent believes, and $\eta r_n P$ is an occurrence of a specifically-used indefinite in the embedded clause of a belief attribution, $c(n)(v)$ will be the specific individual that would have been the speaker's referent of this occurrence of the indefinite in v.

3.6 Speaker as responsible for asymmetry

Our above discussion seems to have led to the following conclusions with respect to intentional identity attributions: When Hob-Nob sentences are uttered out of context, there seems to be an *asymmetry* between what Hob and Nob each have to believe in order to make the attribution true. Nob, but not Hob, has to believe what is attributed both to him *and* to the other, an asymmetry which can be accounted for by any of the *descriptive* approaches towards anaphora. However, when enough context is given, the second agent doesn't have to believe what is attributed to the first agent in order for Hob-Nob sentences to be used appropriately, and for these cases that don't show the asymmetry it is important that pronouns can also be used *referentially*. Thus, asymmetric behaviour should be explained by a descriptive use of pronouns, whereas non-asymmetric behaviour should be explained in terms of referential pronouns.

But this cannot be the whole story. The reason is that Hob-Nob sentences also show asymmetric behaviour whenever the second agent does not believe everything that is attributed to the first agent. Consider the following variant of the original Arsky and Barsky story, again due to Edelberg:

[11] See Stalnaker (1988).

> *Monday*: Smith and Jones have been shot, at opposite ends of Chicago. Arsky and Barsky are investigating both cases, but neither knows that Smith is the mayor or that Jones is the commissioner. Smith and Jones, though hospitalised, are (and are known by both detectives to be) still alive. Arsky and Barsky have discussed the two cases at length, and though they think someone shot Smith and that someone shot Jones, both believe the two cases are entirely unconnected. At this time, neither has anyone in mind as a suspect.
> *Tuesday*: Both Smith and Jones have died of their gunshot wounds. Arsky knows Smith died, and thus now believes that the person who shot Smith murdered him, but doesn't know Jones is dead. Likewise, Barsky knows Jones died, and thus now believes that the person who shot Jones murdered him, but doesn't know Smith is dead. After reflecting on certain similarities between the two cases, Barsky infers that the man who shot Smith is the same person as the man who shot Jones. He communicates this to Arsky, saying, "The man who shot Smith is the man who shot Jones." Arsky disagrees, but Barsky persists in his opinion. (Edelberg, 1986, pp. 16-17)

On Tuesday, (77), repeated as (100), is true and (78), repeated as (101), is false on their most natural readings:

(100) Arsky believes that someone murdered Smith, and
Barsky believes he murdered Jones.

(101) Barsky believes that someone murdered Jones, and
Arsky believes he murdered Smith.

However, this asymmetry cannot be explained by treating the pronouns as abbreviations for descriptions recoverable from the clause in which the indefinite occurs. Barsky does not believe that Smith was murdered, and Arsky does not believe that Jones was murdered. The asymmetry can be explained, however, if we assume that the *speaker*, when he utters (100), has in mind the concept corresponding to 'the one who *shot* Smith' – something that seems plausible if both speaker and hearer are aware of the story above.

3.7 Belief objects and externalism

In the previous sections I have been assuming that the belief states of agents can be modelled by sets of possible worlds. In chapter 2, however, I followed standard dynamic semantics in arguing that presupposition states should be represented in a more complex way, whereby they also contain information about the values of variables/discourse referents. But of course, once we assume that the information associated with discourse referents stands for speaker referents, we might *explain* this information in terms of '*world*'

information, if a reference context is part of a 'world' in the intuitive sense of this word. The value of discourse referent r in possibility $\langle w, c, g\rangle$ is simply the unique speaker's referent of a specific occurrence of an indefinite used in that 'world'. That's why our analysis is so closely related with the E-type approach to anaphora.

Still, our referential analysis of pronouns assumes that the speaker's referent introduced by a specifically-used indefinite is the individual that the speaker had 'in mind'. It is not clear how this intuition should be cashed out if we don't assume that the belief states are structured around *belief objects*. And once we assume that the belief states of *speakers* should be structured around belief objects, there seems to be no reason to deny that the belief states of the *agents* that the belief attributions are about should also contain such belief objects.

Indeed, in van Rooy (2000), I propose, following Edelberg (1992, 1995), that all belief states should be structured around belief objects, and that the concept introduced by a specifically-used indefinite occurring in the embedded clause of a belief attribution corresponds to an object in the belief state of the relevant agent. One way to account for intentional identity attributions is then to assume that the belief objects in the belief states of different agents could be *counterparts* of each other. In van Rooy (1997, 2000), I show that a counterpart theory for the belief objects in information states used for the analysis of intentional identity attributions could be formulated simply as a generalisation of the counterpart theory of objects existing in worlds as formalised in Appendix A to account for *de re* belief attributions.

So, although we don't *have* to assume that the belief states of agents contain belief objects in order to account for Hob-Nob sentences, it's not unreasonable to assume that the concepts introduced by the relevant indefinites *do* correspond to such belief objects.

3.8 Conclusion

Our discussion of intentional identity attributions in this chapter has led to certain conclusions. First, and foremost, to analyze intentional identity attributions we need to take the notion of *speaker's reference* seriously. Second, the problem of intentional identity is not just a problem of *anaphora*; *presuppositions* show the same dependencies. Third, intentional identity attributions are similar to examples of information exchange, but not the same. The difference is due to the fact that it is the *speaker* who is responsible for the anaphoric and presuppositional dependencies, and this difference in responsibility should be implemented by the different kinds of information that main and subordinated contexts contain.

In this chapter we have looked closely at *anaphoric* dependencies across *belief* attributions, and given some attention to similar cases of *presuppositional* dependencies, which will be treated in greater detail in the next chapter. Note that I have neglected intentional identity cases where other attitudes than belief are involved, although the formal

analysis sketched here should also be able to account for dependencies across, for instance, *desire* attributions. These dependencies clearly have a lot in common with the intentional identity attributions that I have concentrated on here, even though there are, I believe, also some important differences. In particular, the assumption of belief objects referred to by means of specifically-used indefinites can help us to analyze anaphoric relations across belief attributions, but does not seem to help in the case of desire attributions. In order to account for anaphoric dependencies in these cases, we must, I will argue, take *belief revision* into account. In chapter 5 I will discuss some analyses of belief revision, and in the last chapter of this book I will show how this can be used in treating the interpretation of some attitude attributions. Before considering belief revision, however, let's first look at the attitude of *presupposition*.

Chapter 4

Presupposition Satisfaction

4.1 Introduction

In traditional pragmatic theories the notion of context plays two roles: (i) it should contain enough information about the conversational situation to determine what is expressed by a sentence; (ii) it should contain enough information about what the participants of the conversation commonly assume about the subject matter of the conversation to determine whether what is said by a speaker is appropriate or not. The central idea behind Stalnakerian pragmatics is that there is a single notion of context that plays both of these two roles, and that both kinds of information modeled by this single context change during a conversation in an interactive way. A context, modeled by a set of possibilities, represents that what is presupposed by the participants in a conversation.

Despite the fact that Discourse Representation Theory (Kamp, Heim) and dynamic semantics (Groenendijk & Stokhof, Veltman) can be looked upon as attempts to incorporate Stalnaker's ideas into a rigorous theoretical model, the resulting dynamic theories differ on some essential points from Stalnaker's suggestions. First, where Stalnaker always argued that the possibilities that are used to represent contexts should be *possible worlds*, proponents of these dynamic theories account for the antecedent-pronoun relation in terms of possibilities that are *finer-grained* than worlds. Second, although Stalnaker always argued for a dynamic view of language use, he didn't give up the traditional distinction between *content* (truth conditions) and *force* (the way a sentence changes a context) of an assertion, while in dynamic semantics the meaning of a sentence is equated with its *context-change potential*. Third, where Stalnaker tried to explain *linguistic presupposition* in terms of what *speakers* normally presuppose by their use of these sentences, and thus taking presupposition to be primarily a *propositional attitude*, dynamic semantics either accounts for presuppositions in a way equivalent to Peter's (1977) three valued logical account (Beaver, 1995, 2001), or (partly) in terms of a syntactic underspecification analysis (van der Sandt, 1992).

The three ways in which standard dynamic semantics differs from Stalnaker's original suggestions are closely related to each other.

First, what is presupposed by the participants in a conversation is according to all a crucial contextual parameter to determine content and appropriateness of sentences. Stalnaker argues that it is an attitude playing a role in action very similar to that of belief. As a result it should be modeled in terms of possibilities whose fine-grainedness is relevant for the analysis of deliberation: possible worlds. Dynamic semanticists – following Lewis (1979a) for the analysis of belief – argue that contexts should consist of possibilities much finer-grained than worlds, i.e., world-assignment pairs. In distinction with Lewis (1979a), however, no-one has ever explicitly argued how this fine-grainedness could be relevant for action. Perhaps because proponents of dynamic semantics have given up the idea that contexts represent that what is presupposed, i.e. a propositional attitude of participants in a conversation. Now, we argued in chapter 2 of this book with Stalnaker (1998b) for a partly referential analysis of anaphoric pronouns, and one of the main reasons for this is that in this way the fine-grainedness of possibilities could, and should, be that of possible worlds. The reason being that on a referential analysis of anaphoric pronouns the use of a singular pronoun comes with a uniqueness assumption.

Second, a three-valued logic accounts for linguistic presupposition in terms of entailment. One of Stalnaker's reasons to account for the behavior of linguistic presuppositions in terms of what speakers presuppose is to be able to account for the intuition many people have that the truth of the linguistic presupposition of a sentence can be *irrelevant* to the truth or falsity of the sentence, or its *content*. All that matters is whether the linguistic presupposition is already satisfied by the context. As a result, we should be able to determine the truth value of sentences in worlds/possibilities *outside* the context; something that is impossible in standard dynamic semantics because no distinction is made between content and force.

Third, if we want to respect the distinction between content and force, we have to be able to determine the truth value of a sentence containing an anaphoric pronoun in possibilities outside the context. It is not at all clear how to do this when no uniqueness requirement is made on the use of singular pronouns. When such a requirement is made, however, it is easy to see that the content-force distinction can be maintained.

In chapter 2 I argued to account for the antecedent-pronoun relation in a way that respects the distinction between content and force, and – by adopting a uniqueness requirement for singular pronouns – model possibilities (essentially) as fine-grained as worlds. In this chapter I will deal with presuppositions. I will think of presupposition more explicitly as a propositional attitude, account for this attitude in possible world semantics, explain some presuppositional phenomena in terms of it, while respecting the distinction between content and force. Before I will do that, however, I will first state the way in which presuppositions are standardly accounted for within dynamic semantics.

4.2 Standard Implementation

According to dynamic semantics, the meaning of a sentence is its *context change potential*, where contexts are identified with information states that represent what is commonly assumed in a conversation. The meaning of a sentence is modeled as an *update function* that takes a context as its argument and has the updated context where the sentence is accepted as its value. Assuming that a sentence cannot be used appropriately in a context that does not entail, or *satisfy*, its triggered presupposition, this function will be *partial*.

I limit myself here, and in most of the rest of this chapter, to the propositional case and represent a context, C, by a set of possible worlds. A possible world is a function from atomic formulae to the two classical truth-values. Just like Veltman (1996), I treat *might*, $\Diamond$, in this section as a test-operator. I will follow Beaver (1995) in using a special presuppositional connective '∂'. We might treat disjunction and implication syncategorematically, by having '$A \vee B$' and '$A \rightarrow B$' stand for '$\neg(\neg A \wedge \neg B)$', and '$\neg(A \wedge \neg B)$' respectively. The update function is defined as follows:

- $[A](C) = \{w \in C |\ w(A) = 1\}$, if A is atomic
- $[\neg A](C) = C - [A](C)$
- $[A \wedge B](C) = [B]([A](C))$
- $[\Diamond A](C) = C$, if $[A](C) \neq \emptyset$, $\emptyset$ otherwise
- $[\partial A](C) = C,$ if $[A](C) = C$, undefined otherwise

The appealing feature of this analysis of presuppositions within dynamic semantics (as stressed by Heim (1983)) is that it seems to solve the projection problem simply by means of rules of interpretation. Assuming that context C satisfies presupposition P iff $[P](C) = C$, we can say that sentence A presupposes P iff for all contexts $C, [A](C)$ is defined only if C satisfies P. As a result, it follows for instance that if A presupposes P, sentences of the form $\neg A$, $\Diamond A$, and $A \wedge B$ do so too, but that $B \wedge A$ need not.

4.3 Presupposition as a propositional attitude

4.3.1 Motivation

An important insight shared by Stalnaker and proponents of dynamic semantics is that we presuppose not only something about the subject matter of conversation, but also about the conversational situation itself. Perhaps the most important kind of information about the conversational situation that agents have presuppositions about is the information that (other) agents presuppose (about the conversational situation). For reasons like this, Stalnaker (1970b, 1973, 1974, 1998b, 2002) argues that presupposition should be thought of as a propositional attitude.

According to the *functional* analysis of attitudes, an agent stands in a certain attitude relation to a proposition, if by means of this relation, together with the assumption that the agent is rational, we can explain the behavior of the agent. Attitudes are seen as dispositional, or functional, states of a rational agent, and these states are individuated by the role they play in determining the behavior of the agent who is in such a state. This picture suggests that contexts represent presuppositions and should also be thought of as propositional attitudes: we have to know what the speaker is presupposing in order to explain his behavior when he is engaged in a conversation.

According to Stalnaker (1970b), we should explain the appropriateness of what someone says not only in terms of what he believes and desires, but also partly in terms of what he presupposes. To be able to explain the actions of rational agents, we must assume that the believers know their own minds, i.e. have *introspective* access to their own minds. In possible world semantics introspective belief states are modeled by an accessibility relation that is *serial*, *transitive*, and *Eucledian*.[1] Stalnaker (1974, 1998b, 2002) has always argued that presupposition should be thought of as a propositional attitude and thus represented in a similar way: by means of an *accessibility relation*. But what do agents presuppose? The standard answer is: that what is common ground between the participants of the conversation. According to discourse representation theory, what is common ground is that what is explicitly represented in a discourse representation structure, a DRS. This DRS, in turn, represents what has been *explicitly* agreed upon by the conversational participants. This suggests that presupposition should by default be fully *introspective*: what is presupposed is also presupposed to be presupposed, and what is not presupposed is also presupposed not to be presupposed.[2] I will represent what is presupposed by a *primitive* accessibility relation R.

Although in the ideal case the context – what is presupposed – represents what is common knowledge for the participants in the conversation, it is clear that ideal conditions do not always obtain. For one thing, what is presupposed need not be true: discourse can be based on an assumption that later turns out to be false. So, presupposition should be represented by an accessibility relation that need *not* be *reflexive*. If presupposition is analyzed in terms of a non-reflexive accessibility relation, we correctly predict that also presupposition is a *non-veridical* propositional attitude.

An important motivation for treating presupposition as a *non-veridical* propositional attitude is that we can then respect the traditional distinction between the *content* and the *force* of a speech act (of assertion), and can separate questions of entailment from

[1] A relation R is *serial* if $\forall x : \exists y\ xRy$; *transitive* if $\forall x, y, z :\ (xRy\ \&\ yRz) \rightarrow xRz$; and *Euclidean* if $\forall x, y, z : (xRy\ \&\ xRz) \rightarrow yRz$.

[2] See also Fernando (1997) for an analysis of context where full introspection is assumed. Stalnaker (2002) suggests that what is presupposed by an agent is that what she believes is commonly believed by the discourse participants. This has as a result, however, that the attitude of presupposition does not obey negative introspection, because more things can be taken to be commonly believed than what is explicitly agreed upon.

questions of presupposition satisfaction. The views that we should separate content from force, and valuation of truth from presupposition satisfaction are closely related, and they have been defended consistently over the years by Stalnaker.

> I suggested that an assertion should be understood as a proposal to change the context by adding the content to the information presupposed. [...] Meaning determines the content of an assertion as a function of the context, and the assertion rule takes the prior context set to a posterior context set [...] Some of the dynamic semantic theories subsequently developed by linguists have blurred the distinction between content and force by combining the two steps (meaning plus prior context to content, and prior context plus content to posterior context) into one. [...] I think this streamlined representation captures much of what is important about the dynamic process of speech, but what it leaves out is the possibility of evaluating the truth or falsity of what is said relative to possible situations that are not compatible with the prior context. Sometimes when a statement rests on false presuppositions, the question of the actual truth of the statement does not arise, but other times a speaker may succeed in making a claim that is actually true or false, even when taking for granted, in making the claim, something that is in fact false. In such cases, our semantic theory should tell us what is said, and not just how what is said changes the context. Sentences that say different things in some contexts may nevertheless change contexts in the same way. (Stalnaker, 1999, p. 11)

One example he discusses for which the separation of content from force seems crucial is Donnellan's (1966) case of the referential use of definite descriptions. When the description in the sentence *The man drinking a martini is a philosopher* is used referentially, the proposition intended to be communicated/expressed might be true, although it presupposes a falsehood. Stalnaker (1973, 1974), followed by linguists like Karttunen and Peters (1979), argues that in general the proposition expressed by a sentence might be true independent of the truth of the presupposition: a sentence like *Even Bill likes Mary* can be true without it being unlikely that Bill likes Mary. So, although the use of the sentence gives rise to this presupposition, the sentence by itself doesn't entail it. This suggests that the truth value of, or proposition expressed by, a sentence should be determinable independently of the truth of the presupposition. Note, however, that this is impossible in standard dynamic semantics, where all of the attention is given to the *update* of what is presupposed, and where *truth* is treated as at best a *derived* notion. But if in semantics *truth* and *truth conditions* are of primary importance, we should be able to say when a sentence is true, even if it is interpreted with respect to a presupposition state that is non-veridical. Thus, truth and presupposition satisfaction should be accounted for on *different dimensions*. To be able to do this, we have to define the truth conditions of sentences in a manner different from that of standard dynamic semantics – namely, one

that is more independent of the update function of what is presupposed.

The non-veridicality of contexts suggests that we should treat the valuation of truth separately from context change – distinguish content from force. In the remainder of this chapter I will show how we can systematically account for presupposition satisfaction without giving up the possibility of determining the content of a sentence separately from the way it changes the context. For context change, I will rely mainly on work in dynamic epistemic semantics, where updates are defined in terms of *eliminating arrows* instead of eliminating worlds.[3]

4.3.2 Formalization

When a speaker presupposes something, he presupposes it in a world or a possibility. A possibility will be represented by a {*pointed model*, $\langle R, w\rangle$},[4] where w is a distinguished world representing the actual world and should be thought of as a valuation function from atomic propositions to truth values and where R is the presuppositional accessibility relation that is (by default) *serial*, *transitive* and *Euclidean*. I will take $R(v)$ to be the worlds accessible from v: $\{u \in W : vRu\}$. As a result, it will be the case that what is presupposed is introspective: $\forall v, w$: if $v \in R(w)$, then $R(v) = R(w)$, although it need not be veridical, i.e., it might be that $w \notin R(w)$. To determine in possibility $\langle R, w\rangle$ whether P is presupposed, we have to check what is presupposed in this possibility, $R(w)$.

The two-dimensional (or four-valued) analysis of presupposition that was popular in the seventies treats the logic of truth and that of presupposition at separate dimensions. This is appealing because sometimes a sentence can, intuitively, be true, although its presupposition is false. Standard dynamic semantics treats conjunction in an *asymmetric* way: the second conjunct should be interpreted with respect to the initial context updated with the first conjunct. This is a desirable feature of a framework to account for the asymmetric behavior of presuppositions in conjunctive sentences. In this section I will combine the desirable features of both the two-dimensional and the dynamic analysis of presuppositions. Thinking of presupposition as a non-veridical propositional attitude, we can account for the dynamic aspects of presupposition satisfaction without giving up the idea behind a two-dimensional analysis of presupposition satisfaction. That is, although we will predict that conjunction behaves *asymmetrically* with respect to presupposition satisfaction, '*and*' will still be treated in a *symmetric* way. The reason is that truth and presupposition satisfaction are defined *separately* from the update function (although they

[3] Updating through the elimination of arrows instead of worlds has been used, among others, by Veltman (1996). Its limitations for multi-agent settings are discussed in Gerbrandy (1999).

[4] In fact, a pointed model is a tuple like $\langle M, w\rangle$, where M is a modal model $\langle W, R_1 ... R_n, I\rangle$. I will assume that the set of worlds W of all pointed models is the same, take w to be an interpretation function for atomic sentences, and concentrate only on one accessibility relation. If we think of a world as representing everything that is the case, including some modal facts, a pointed model should be thought of as such a world.

will be defined simultaneously). Making use of Beaver's (1995) presupposition operator, I will represent an atomic sentence A that presupposes P as follows: $\partial P \wedge A$. For the time being, I will concentrate only on the truth-conditional connectives. I will assume that a sentence has two values: (i) a sentence is true or false, i.e. 1 or 0; (ii) a sentence has no presupposition failure or it has one, i.e. + or -. The combined *truth* and *presupposition satisfaction conditions* of sentences are given below (where '$\cdot$' is a placeholder):[5]

- $[[A]]^{R,w} = \langle 1/0, + \rangle$, iff $w(A) = 1/0$, if A is atomic (then always defined)
- $[[\neg A]]^{R,w} = \langle 1/0, + \rangle$ iff $[[A]]^{R,w} = \langle 0/1, + \rangle$, $\langle \cdot, - \rangle$ otherwise
- $[[A \wedge B]]^{R,w} = \langle 1, + \rangle$ iff $[[A]]^{R,w} = \langle 1, + \rangle$ and $[[B]]^{Upd(A,R),w} = \langle 1, + \rangle$
  $= \langle \cdot, - \rangle$ iff $[[A]]^{R,w} = \langle \cdot, - \rangle$ or $[[B]]^{Upd(A,R),w} = \langle \cdot, - \rangle$
  $= \langle 0, + \rangle$ otherwise
- $[[\partial A]]^{R,w} = \langle 1, + \rangle$ iff $\forall v \in R(w) : [[A]]^{R,v} = \langle 1, + \rangle$
  $= \langle \cdot, - \rangle$ otherwise

Observe again that the presupposition value of a conjunction is determined in a *symmetric* way. That is, if either A or B has a presupposition failure, the conjunction $A \wedge B$ will have a presupposition failure as well. However, to determine the presupposition value of a conjunction of the form $A \wedge B$ in possibility $\langle R, w \rangle$, we look at the presupposition value of B in possibility $\langle Upd(A, R), w \rangle$ – the update function is being relevant here. This is the point at which we take over the insights of dynamic semantics. The update $Upd(A, R)$ is defined as follows:

- $Upd(A, R) = \{\langle u, v \rangle \in R |\ [[A]]^{R,v} = \langle 1, + \rangle\}$.

Notice that this update function is *eliminative*, but instead of eliminating worlds in $R(w)$ it eliminate tuples, or *arrows*, in R. It eliminates all arrows in R that point to an non-A-world. This has the effect that after the update of R with A, not only all worlds v accessible from w verify A, but also all worlds u accessible from v make A true. Thus, after the update with A it is not only presupposed that A, but it is also presupposed to be presupposed that A. Moreover, on the assumption that R is fully introspective, $Upd(A, R)$ will be fully introspective as well. Also after the update, everything that is not presupposed is also presupposed to be not presupposed.

Our analysis is very similar to standard dynamic semantics. If we would say that $[[\Diamond A]]^{R,w} = \langle 1, \cdot \rangle$ iff $\exists v \in R(w) : [[A]]^{R,v} = \langle 1, \cdot \rangle$ and assume that possibility statements don't have any dynamic effect,[6] we predict just like Veltman (1996) an asymmetry between

[5] Although I use a four-dimensional logic, I am not explicit about when a sentence is true or false, although its presupposition is not satisfied. But this is needed if we want to allow *Even John was there* to be true although it is not presupposed that John's being there was unlikely (thanks to Kai von Fintel for reminding this to me). However, there is no principle problem of distinguishing those cases as well.

[6] Though we will give a somewhat different analysis of possibility statements later.

$\Diamond A \wedge \neg A$ and $\neg A \wedge \Diamond A$; the former is okay, the latter is not. However, this contrast in acceptability is explained in a somewhat difference way: Veltman's explanation appeals to acceptability of update, while we explain the contrast in terms of truth. We predict that the former sequence can be true, but the latter cannot.

If we assume that sentence A presupposes P iff $\forall\langle R, w\rangle$: if $[[A]]^{R,w} = \langle \cdot, + \rangle$, then $\forall v \in R(w)$: $[[P]]^{R,v} = \langle 1, + \rangle$, the above implementation gives rise to the same presuppositional predictions as the standard implementation of the satisfaction account. In particular, on the assumption that *John stopped smoking* gives rise to the presupposition that John used to smoke, this implementation predicts that sentences like *John didn't stop smoking* and *John stopped smoking and Mary is sick* will also gives rise to this presupposition, but *John used to smoke and he stopped doing so* will never give rise to presupposition failure.

Although the predictions of the above implementation of the satisfaction approach are similar to the predictions on the standard approach, there are still some important differences. First, note that by treating presupposition as a propositional attitude, we can evaluate in a *distributive* way whether a presupposition associated with a sentence is satisfied by what the speaker presupposes. This is possible, of course, because we have represented in a single possibility all the information that is normally represented only in a whole context/information state. Second, and related, we can now account for the dominant view in the seventies that presupposition satisfaction and truth should be evaluated at *different dimensions*.

> [...] if presupposition is defined independently of truth-conditions, then we can separate the question of entailment relations from the question of presupposition. [...] one may say that sometimes when a presupposition is required by the making of the statement, what is presupposed is also entailed, and sometimes it is not. One can say that "Sam realizes that P" entails that P –the claim is false unless P is true. "Sam does not realize that P," however, does not entail that P. That proposition may be true even when P is false. All this is compatible with the claim that one is required to presuppose that P whenever one asserts or denies that Sam realizes it. (Stalnaker, 1974, p. 54)

We have already seen that according to Karttunen & Peters (1979) and others a sentence like *Even Bill likes Mary* presupposes something that it does not entail. Thus, the sentence can be true without it actually being unlikely that Bill likes Mary, because what is presupposed need not be true. Notice that we can now account for this intuition without assuming with Karttunen & Peters (1979) that we should thus *represent* presuppositions separately from assertions. On the other hand, we can also account for the intuition that a factive verb both presupposes and entails that its complement is true.[7] To analyze *Sam realizes that P* we add the following construction to the language: if P is a sentence,

[7] Throughout the paper I will assume the same for an aspectual verb like *stop*.

$Real(s, P)$ is a sentence too. To interpret the formula, we add a primitive reflexive accessibility relation to the model, K_s, modeling what Sam realizes.[8] The formula is then interpreted as follows:

- $[[Real(s,P)]]^{R,w} = \langle 1,+\rangle$ iff $\forall v \in K_s(w) : [[P]]^{R,v} = \langle 1,+\rangle$
  $= \langle 0,+\rangle$ iff $\exists v \in K_s(w) : [[P]]^{R,v} = \langle 0,+\rangle$, $\langle \cdot,-\rangle$ otherwise

Notice that because K_s is reflexive, according to this analysis the formula entails, but does not presuppose, that P. To account for the presupposition, we represent the sentence *Sam realizes that P* by the following formula $\partial P \wedge Real(s, P)$, which both presupposes and entails that P. If we now represent *Sam does not realize that P* by $\neg(\partial P \wedge Real(s, P))$, this sentence presupposes that P, but can still be true in case P is false (in case $w \notin R(w)$).

4.4 Quantification and anaphora

4.4.1 The binding problem

The traditional problem for a two-dimensional analysis of presuppositions is Karttunen & Peters's (1979) *binding problem*: their false prediction that the individual that satisfies the presupposition of a sentence like *Someone managed to succeed George V on the throne of England* need not be the one who makes the sentence true by actually having succeeded George V. Here I show that this problem will not arise in our framework if we extend it to the predicate-logical case.

The binding problem of Karttunen & Peters' (1979) was due to the fact that they *represented* presupposition and assertion separately. Our analysis, instead, only *interprets* them at different dimensions. Let us assume, for simplicity, that indefinites are analysed simply as existential quantifiers and that bound variables are interpreted with respect to an additional assignment function. Then, we can interpret a sentence of the form $\exists xA$ in $\langle R, w, g\rangle$ as follows:

- $[[\exists xA]]^{R,w,g} = \langle 1,+\rangle$ iff $\exists d \in D : [[A]]^{R,w,g[x/d]} = \langle 1,+\rangle$
  $= \langle 0,+\rangle$ iff $\exists d \in D : [[A]]^{R,w,g[x/d]} = \langle \cdot,+\rangle$ and
  $\forall d \in D$: if $[[A]]^{R,w,g[x/d]} = \langle \cdot,+\rangle$, then $[[A]]^{R,w,g[x/d]} = \langle 0,+\rangle$,
  $= \langle \cdot,-\rangle$ otherwise

Now we represent K&P's problematic sentence abstractly as follows: $\exists x[\partial Px \wedge Qx]$. An easy calculation shows that this formula is predicted to be true and appropriate in $\langle R, w, g\rangle$ just in case $\exists d \in D : \forall v \in R(w) : [[Px]]^{R,w,g[x/d]} = \langle 1,+\rangle$ & $[[Qx]]^{R,w,g[x/d]} = \langle 1,+\rangle$. Thus, it is required that the same individual has to satisfy both the presuppositional

[8] Our simple update function has limitations here: if we would attribute to Sam attitudes about what the discourse participants presuppose, things go wrong. I will ignore such attributions in this paper. See, among others, Gerbrandy (1999) for an analysis in which this problem is overcome.

part and the assertive part: the binding problem does not occur. Notice that if we assume that $\forall xA$ is an abbreviation for $\neg\exists\neg A$, we predict that a sentence like *every man loves his wife* can only be true and appropriate if (i) there is a man who has a wife, and (ii) every man who has a wife loves her. Thus, the prediction is almost identical to that of van der Sandt (1992), although we don't need to make use of something like *intermediate accommodation*.

4.4.2 Anaphora

In the two previous chapters we have seen that to account for anaphoric dependencies across the sentential boundary, we have to assume that possibilities have to contain more information than possible worlds as they are standardly conceived of. In particular, the possibilities should contain information about speaker's reference of occurrences of specifically used indefinites, and should be able to change to account for anaphoric dependencies. Thus, to take indefinites and pronouns into account, we have to make our accessibility relation one between more fine-grained possibilities. Let us assume that the set of possibilities I is a set of functions from (i) n-ary predicates to their interpretations; (ii) reference functions to individuals; and (iii) discourse referents to individuals. If $\eta r_n P$ is interpreted in possibility $i = \langle w, c, g\rangle$, then $i(n) = c(n)$ is the speaker's reference of the occurrence of the indefinite in i, and the dynamic effect will be that from now on discourse referent r will be assigned to $i(r)$ in i, i.e. $i(r) = i(n)$. Let us define $R[^x/_r]$ as $\{\langle i[^r/_{i(n)}], j[^r/_{j(n)}]\rangle : \langle i, j\rangle \in R\}$. Now we define the update of R with $\eta r_n P$, $Upd(\eta r_n P, R)$, as $R[^r/_n]$.

Thus, just as in chapter 2 we assumed that in the actual possibility the speaker's reference of the indefinite is introduced (although this need not be an individual that makes the sentence true), now we assume that this also happens in each possibility that is compatible with what is presupposed: the speaker's reference of the indefinite in that possibility is introduced. But there should be a difference between the actual possibility and the ones compatible with what is presupposed: in the latter, the speaker's reference should also verify the sentence.

Now we have to know how pointed models should be updated. We (tentatively) propose the following (forgetting about descriptive pronouns):

- $Upd(t_1, ..., t_n, \langle R, i\rangle) = \langle\{\langle j, k\rangle \in Upd(t_n, ..., Upd(t_1, R)...) | \langle [[t_1]]^k, ..., [[t_n]]^k\rangle \in I_k(P)\}, Upd(t_n, ..., Upd(t_1, i)...)\rangle$
- $Upd(\neg A, \langle R, i\rangle) = \langle Upd(\neg A, R), i\rangle$
- $Upd(A \wedge B, \langle R, i\rangle) = Upd(B, Upd(A, \langle R, i\rangle))$

The (rigid) *truth* and *presupposition satisfaction conditions* of the new clauses are defined just as in chapter 2 and as earlier in this chapter. In this way it follows – as we argued for in section 2.5 – that not only the actual possibility but also all possibilities

consistent with what is presupposed presuppose something about the speaker's referent of every specifically used indefinite.[9]

In the remainder of this chapter I will limit myself again to the propositional case.

4.5 No cancellation or local accommodation

Consider the following well known problematic examples for the traditional satisfaction theory:

(102) Frank doesn't *know* that the earth is flat, because the earth isn't flat.

(103) It is *possible* that John used to smoke,
and it is possible that he just *stopped* doing so.

(104) Either John *stopped* smoking, or he just *started* doing so.

These example are problematic for the standard satisfaction approach because this account wrongly predicts in all these cases that the sentences give rise to presuppositional readings that intuitively are not the case. Sentence (102) is predicted to presuppose that the earth is flat, which is in conflict with what is asserted; sentence (103) is falsely predicted to presuppose something what is only claimed to be possible in the first conjunct; and (104) is predicted to give rise to two mutually incompatible presuppositions, which is absurd.

Traditionally, these examples gave rise to the hypothesis that presuppositions can sometimes be *cancelled* for reasons of informativity,[10] while Heim (1983) and van der Sandt (1992) argue that presuppositions should sometimes be *locally accommodated*. But there are problems with both proposals, both formally and conceptually. The *formal* problem is that it is not at all clear how to account for cancellation and/or local accommodation in the framework of the satisfaction approach. The *conceptual* problem for cancellation is that it becomes unclear why the presupposition trigger was used in the first place, and for local accommodation how to explain what is supposed to be going on when we locally accommodate a presupposition.

In the rest of this chapter I will suggest how to account for these apparent counterexamples of the satisfaction analysis by assuming that there might be more than one information state around that could satisfy the triggered presupposition. I will do this all in terms of the above stated possible world analysis.

[9]It is interesting to see that if we assume that (i) *managed* to do x presupposes that it was difficult to do x and (ii) the indefinite in *Someone managed to succeed George V on the throne of England* is used specifically, we predict a somewhat *weaker* presuppositional reading than discussed in section 4.1. Acccording to this alternative analysis we don't need to have a *de re* presupposition about a particular person. I am not sure whether this is a better prediction than what we discussed above.

[10]We might interpret the proposals of Gazdar (1979), Soames (1982), van der Sandt (1988), and some remarks of Stalnaker (1974) in this way.

4.5.1 Denials

I assumed above that what is presupposed by individual a should be represented by a primitive accessibility relation R_a. Although in the ideal case the other agent, b, of the same conversation presupposes the same, this need not really be the case. Thus, in a conversation between a and b there might be several presupposition states around. Because it is the *speaker* who is responsible for what she says, the presuppositions of what she says should in the first place be satisfied with respect to her own presupposition state. This, at least, is normally the case. However, so I want to argue, this is not so for sentences like:

(102) Frank doesn't *know* that the earth is flat, because the earth isn't flat.

This example is problematic for the standard satisfaction approach, because it both presupposes that the earth is flat, and asserts that the earth is not flat. If truth and presupposition satisfaction should be analyzed with respect to the same context, how should we account for such examples? The standard answers, as we have seen, are *cancellation* and *local accommodation*. In this section I want to suggest that (102) does indeed give rise to the presupposition that the earth is flat, as the standard satisfaction account predicts, but that it is not the presupposition of the *speaker*, but rather that of the *addressee*. The idea is that a sentence like (102) is typically uttered after the other participant in the conversation has made clear (perhaps, but not necessary, by an explicit claim) that he presupposes that the earth is flat. In terms of van der Sandt (1991), this means that (102) is typically used as a *denial*. To account for the intuition that sometimes a speaker indicates that a presupposition is made not by himself, but rather by the other participant in a conversation, we will index the presupposition operator by the relevant agent. Thus, $\partial_j P$ will mean that agent j presupposes that P is the case. As might be expected, we will represent (102) by the following formula:

(105) $\neg(\partial_j P \wedge know(f, P)) \wedge \neg P$

Assuming that K_f is the reflexive accessibility relation that models what Frank knows, the first conjunct is analyzed as follows:

- $[[\neg(\partial_j P \wedge know(f,\ P))]]^{M,w} = \langle 1, + \rangle$ iff $[[\partial_j P \wedge know(f, P)]]^{M,w} = \langle 0, + \rangle$ iff
  $\forall v \in R_j(w):\ [[P]]^{M,v} = 1$ and $[[know(f, P)]]^{M,w} = 0$ iff
  $\forall v \in R_j(w):\ [[P]]^{M,v} = 1$ and $\exists u \in K_f(w):\ [[P]]^{M,u} = 0$

Notice that this first conjunct is obviously compatible with the second one: if, and only if, w doesn't make P true and $w \notin R_j(w)$ both conjuncts can be true. The point of what the speaker says by (102) is in fact that the other participant presupposes something that is false: the sentence as a whole can be true only in case R_j is non-reflexive. Thus, we might say, the sentence is used as a presuppositional denial. We can conclude that to account for denials we don't have to assume that presuppositions are cancelled or locally accommodated: they have to be satisfied, but not necessarily by the information state that represents what the *speaker* presupposes.

4.5.2 Modal subordination

According to standard dynamic semantics (Veltman 1996), the embedded sentence of 'possibly A' should be interpreted with respect to the same context as the whole sentence. This gives rise to the prediction that 'possibly A' triggers the same presupposition as A itself. However, if it has already been established that it is possible that John used to smoke, i.e. after (106a) has been asserted, (106b) need not presuppose that John used to smoke.

(106) a. It is *possible* that John used to smoke,

b. and it is *possible* that he just *stopped* doing so.

The phenomenon that a modal expression depends for its interpretation on another modal, as illustrated by (106a)-(106b), is via Roberts (1989) known as 'modal subordination'.

There are by now many interesting analyses of modal subordination around (e.g. Roberts, 1989; Kibble, 1994; Geurts, 1998; Frank, 1997), but none of them seems to be compatible with the view that we should represent what is presupposed in terms of a *single* accessibility relation. In this section I want to show how this is possible for possibility statements, leaving other cases to van Rooij (to appear).[11]

The basic idea is very simple: possibility statements introduce an *ordering* on the worlds. However, because we assume that what is presupposed is a propositional attitude and should be represented by an accessibility relation, we can implement this idea in an appealing way. Following Veltman's (1996) analysis of *normally*, I will assume that the dynamic effect of a possibility statement is that the worlds that make the embedded clause true are the most preferred worlds by eliminating arrows from A-worlds to $\neg A$-worlds.

- $Upd(\Diamond A, R) \;=\; \{\langle u, v\rangle \in R |$ if $[[A]]^{R,u} = \langle 1, +\rangle$, then $[[A]]^{R,v} = \langle 1, +\rangle\}$

According to the update function, possibility statements *disconnect* A-worlds from $\neg A$-worlds, although A-worlds can still be seen from $\neg A$-worlds and from actual world w. Suppose that before the update $R = \{\langle w, v\rangle, \langle w, u\rangle, \langle v, v\rangle, \langle v, u\rangle, \langle u, u\rangle, \langle u, v\rangle\}$, where v is an A-world and w and u are $\neg A$-worlds. Then R is introspective: $R(w) = R(v) = R(u) = \{v, u\}$. After the update with $\Diamond A$, however, the new accessibility relation $Upd(\Diamond A, R)$ won't be introspective anymore: the tuple $\langle v, u\rangle$ will be eliminated, which means that $Upd(\Diamond A, R)(w) \neq Upd(\Diamond A, R)(v) = \{v\} \neq Upd(\Diamond A, R)(u)$.[12] Thus, if R was Euclidean before the update with $\Diamond A$, it won't be Euclidean anymore afterwards.

[11] In this paper I also discuss the alternative approaches to modal subordination.

[12] This update rule is defined on the assumption that either $w \notin R(w)$ or w is not an A-world, because otherwise we would falsely predict that after the use of the possibility statement only other A-worlds would be accessible from w. In general we cannot make this assumption, of course. Fortunately, there are technical ways to solve this problem. One way is to assume that w is the distinguished actual world, and that we change the update rule for possibility statements as follows: $Upd(\Diamond A, R) = \{\langle u, v\rangle \in R |$ (if $[[A]]^{R,u} = \langle 1, +\rangle$, then $[[A]]^{R,v} = \langle 1, +\rangle$) or $u = w\}$. Because our technical problem has a simple solution, I will ignore this complication in the main text.

Possibility statements will be interpreted as follows:

- $[[\Diamond A]]^{R,w} = \langle 1, + \rangle$ iff $\exists v \in R(w) : [[A]]^{R,v} = \langle 1, + \rangle$
  $= \langle 0, + \rangle$ iff $\exists v \in R(w) : [[A]]^{R,v} = \langle \cdot, + \rangle$ and
  $\forall v \in R(w)$: if $[[A]]^{R,v} = \langle \cdot, + \rangle$, then $[[A]]^{R,v} = \langle 0, + \rangle$,
  $= \langle \cdot, - \rangle$ otherwise

According to this rule it holds that if A presupposes P, $\Diamond A$ can be used appropriately only if it is assumed to be possible that P is presupposed. Because out of context (or so we assumed) it holds that $\forall v \in R(w) : R(v) = R(w)$, under normal circumstances $\Diamond A$ presupposes the same as A itself. However, it also can account for the sequence (106a)-(106b), where the presupposition of the embedded clause of (106b) is not a presupposition of its embedding sentence as a whole. The reason is that after the interpretation/update of (106a) there is a world v consistent with what is presupposed in the actual world w in which John used to smoke and in which it is presupposed that John used to smoke. Thus, because from such a world v only worlds are accessible in which John used to smoke, the embedded sentence of (106b) can be interpreted appropriately as well.

The concrete accessibility relation R discussed above illustrates what it means that after the update of R with $\Diamond A$, the A-worlds are the *preferred* ones: although in each $v \in R(w)$ it was the case that both $\Diamond A$ and $\Diamond \neg A$ were true, this is only the case for $\Diamond A$ for all $v \in Upd(\Diamond A, R)(w)$.

Notice that if we take $\Box A$ to be an abbreviation of $\neg \Diamond \neg A$, we predict that A has to be interpreted only in possibilities that satisfy the presupposition of A: $\Box A$ has value $\langle 1, + \rangle$ in $\langle R, w \rangle$ iff $\exists v \in R(w) : [[A]]^{R,v} = \langle \cdot, + \rangle$ and $\forall v \in R(w)$: if $[[A]]^{R,v} = \langle \cdot, + \rangle$, then $[[A]]^{R,v} = \langle 1, + \rangle$. But this means that $\Box(\partial P \wedge A)$ can only be true in $\langle R, w \rangle$ iff either P itself is presupposed and A is true in all accessible worlds, or it is presupposed that P is possible and A is true in all accessible P-worlds.

In this section I have only sketched how a modal analysis of modal subordination can also account for a traditional problem of the satisfaction approach towards presuppositions. In van Rooij (to appear), this analysis is extended such that it also can account for other traditional problems involving disjunctive, conditional and negative sentences, and attitude attributions.

4.6 Conclusion

In this chapter I have argued that presupposition should be thought of as a propositional attitude. This allows us to combine the strong points of the two-dimensional analysis popular in the seventies and the standard dynamic analysis of the late eighties and nineties. As for the former, presupposition satisfaction is determined (almost) independently from truth. This enables us to separate the question of entailment relations from the question

of presupposition. Our analysis differs from other two-dimensional analyses of presupposition proposed in the seventies because it doesn't have to assume that the assertive and presuppositional part of an utterance should be represented at separate representations. I suggested that this helps us to solve the binding problem. Our analysis shares the most appealing feature of the dynamic analysis as well: what is presupposed by a sentence follows from the interpretation rules.

Furthermore, I have suggested how our analysis might account for some (apparent) counterexamples of the satisfaction approach by assuming that there might be more than one information state around that could satisfy the triggered presupposition. This holds both for denials and for modal subordination.

Chapter 5

Conditionals and belief change

5.1 Introduction

In this chapter I discuss conditionals, and the analysis of belief change. There are several reasons for this. In chapter 1 I argued that belief attributions are highly context dependent, and in the following chapters I suggested how the context might change during the interpretation of a discourse. The main argument of this chapter will be that what is expressed by a conditional depends crucially on context, and that some apparent puzzles for the interpretation of conditionals can be solved when we take context change seriously.

I will argue that all conditionals state propositions and should be analyzed uniformly, but what proposition is expressed by a conditional is context dependent, depending on the beliefs, intentions, and presuppositions of language users. In particular, for the analysis of indicative conditionals, the proposition expressed might depend on the *belief* state of the speaker. Just as in earlier chapters, to account for certain problems that arise for this hypotheses, in particular Gibbard's problem, we use diagonalisation to determine what proposition is expressed by certain conditionals. The analysis of counterfactuals depends on context too, and on the way the context changes during a conversation. This might suggest that all the complexities of counterfactual conditionals are due to the complicated ways the interpretation of the counterfactual depends on context. Indeed, it might be argued that as a result conditionals behave *semantically* as strict conditionals, but that the accessibility relation in terms of which the strict conditional is analyzed varies heavily with context. I argue that this can be worked out, and that indeed many of the traditional problems for the strict conditional account can be explained away, but that, perhaps, it would have to make use of accommodation in a too heavy way. At the end of this chapter I will also show how the traditional Lewis/Stalnaker analysis can account for certain puzzles, by showing how the selection function can change its denotation during a discourse.

In this chapter it will be argued that the analysis of conditionals should be closely related with the way agents would change their beliefs on learning new information. In a large part of this chapter I will discuss the question how to analyze *belief change*, and how to represent the belief revision policies of agents. This gives rise to a richer representation

of belief states than is commonly assumed, and I will argue in the next chapter that this richer representation is important for the analysis of, among others, evidential and buletic attitude attributions.

The remainder of this chapter is organised as follows. I first discuss the familiar analyses of counterfactuals by Lewis and Stalnaker. Next, I go into Stalnaker's proposal to make a connection between the Ramsey test idea underlying his own analysis of conditionals, and the Bayesian account of rational conditional belief. Lewis' triviality result showed, however, that the most straightforward way to make this connection won't work. After a discussion of Lewis' result, I will give in the following section an overview of the several attempts to escape the threat of triviality. I will argue that the best way to escape triviality is to assume that the proposition expressed by a conditional sentence is very context dependent. At the end of this chapter I will use this idea to account for certain puzzles that arise if we assume that conditionals are to be analyzed with respect to a selection function, or an accessibility relation. First, Gibbard's puzzle is discussed, and next the puzzle of how to analyze the inappropriateness of certain sequences of counterfactual conditionals.

5.2 The Lewis/Stalnaker analysis of conditionals

Let us assume that sentences of the form *if A then B* should have a uniform analysis. Then the question arises of what the right analysis could be. The conditional *If the butler didn't do it, the gardener did* can be denied without denying the butler's guilt. That's why the material implication isn't a good representation of natural language *if A then B*. In this respect, Belnap's (1970) analysis of conditional assertion is better. It analyses the assertion of a conditional as the assertion of the consequent conditionally on the truth of its antecedent. But because the antecedent of a counterfactual is assumed to be false, this analysis cannot account for an important group of conditionals. It seems we have to consider other ways the world might have been to interpret these conditionals. In other words, we have to go from an extensional to a modal analysis of conditionals. In this way we can analyze the conditional *if A then B*, as follows:

If A and .... then B

But by analyzing conditionals as strict conditionals, the blank can only be filled by all maximal sets of (invisible) premises that are consistent with A. Because both *If I had shirked my duty, no harm would have ensued* and *If I had shirked my duty and you had too, harm would have ensued* can be true simultaneously, the strict conditional analysis will not do.[1] Assuming a fixed accessibility relation for the necessity operator of the strict conditional, it not only cannot account for the non-monotonic behaviour of counterfactuals

[1]Stalnaker and especially Lewis argued that the strict conditional account can also not account for evidence against the validity of contraposition and transitivity. Most authors agree with Stalnaker and Lewis, but we will come back to this in section 5.12.

and conditional assertions, it also does not make the interpretation of the conditional flexible enough to make it dependent on what the speaker had in mind. The actual world, the influence of the intention of the speaker and the non-monotonicity of conditionals must be reflected somehow in the right analysis of conditionals. Lewis and Stalnaker developed an analysis of conditionals[2] in which those three requirements are met by introducing (context dependent) *selection functions* that take as arguments both the actual world and the antecedent of the conditional. In this way, the blank can be filled with enough but not too many additional premises. We will turn now to their possible world analysis.

In possible world semantics, a proposition expressed by a sentence is equated with the set of possible worlds in which the sentence is true.[3] If we represent conditionals as $A > C$, Stalnaker (1968) and Lewis (1973) gave the following analysis:

$A > C$ is true at w if some $A \wedge C$-worlds are closer to w than any $A \wedge \neg C$-worlds.

To make sense of this definition it is necessary to explain what 'closer A-world to w than' means. Therefore an ordering relation on the accessible worlds (but let us assume that all worlds are accessible) with respect to w, $\leq_w$, is defined which meets the following conditions:

(a) reflexivity; $\quad \forall w' : w' \leq_w w'$;

(b) transitivity; $\quad \forall w', w'', w''' : (w' \leq_w w'' \;\&\; w'' \leq_w w''') \Rightarrow w' \leq_w w'''$;

(c) connectedness; $\quad \forall w', w'' : w' \leq_w w''$ or $w'' \leq_w w'$;

(d) strong centering; $\quad \forall w' : w \neq w' \Rightarrow (w \leq_w w'$ and $w' \not\leq_w w)$.

This ordering relation is reflexive, transitive, connected, and obeys the strong centering assumption. So, the relation $\leq_w$ is a weak ordering that meets the strong centering assumption. The intuitive meaning of $w' \leq_w w''$ is that w' is at least as close (or similar) to w as w'' is.

The relation of similarity leads to a notion of *sphere around* w. A sphere around w is a set of possible worlds. Two possible worlds w', w'' are in the same sphere around w iff $w' \leq_w w''$ and $w'' \leq_w w'$. Note that by the strong centering assumption $\{w\}$ is a distinguished sphere around w. The spheres can be ordered by the following $\Re_w$-relation:

$$\forall X \subseteq W, Y \subseteq W, \quad \Re_w(X, Y) \quad \text{iff} \quad w' \leq_w w'' \text{ for all } w' \in X, w'' \in Y$$

This $\Re_w$-relation totally orders the spheres relative to w. A system of spheres centred on $\{w\}$ is a collection $\mathbf{S}$ of subsets of W, $\{S_i : i \in \mathbf{N}\}$, that satisfies the following conditions:

[2]Lewis claims that the analysis only works for counterfactuals. Stalnaker argued that the analysis also applies to indicative conditionals.

[3]In this chapter we will use capitals to denote both natural language sentences and the propositions they express. I hope this will never lead to confusion.

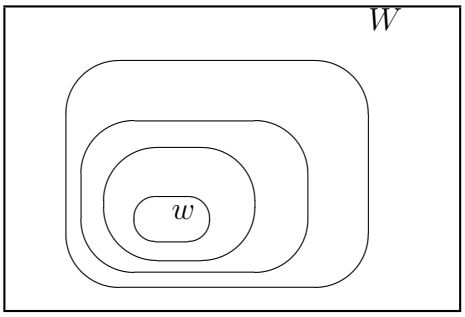
($S1$) $\mathbf{S}$ is totally ordered by $\subseteq$; $(S_i, S_j \in \mathbf{S}\ \&\ i \leq j) \Rightarrow S_i \subseteq S_j$;

($S2$) $\{w\}$ is the $\subseteq$-minimum of $\mathbf{S}$; $\{w\}, S \in \mathbf{S} \Rightarrow \{w\} \subseteq S$;

($S3$) $W \in \mathbf{S}$, and

($S4$) If there is a sphere in $\mathbf{S}$ intersecting a proposition A, there is a smallest sphere in $\mathbf{S}$ intersecting A.

Let's call the system of spheres induced by this ordering relation $\mathbf{S}$. This ordering relation can be pictured as follows:



Let's call the smallest sphere $\mathbf{S}$ that has a non-empty intersection with A, S_A. The set of the most similar A-worlds to w is now $A \cap S_A$.

By using only the above constraints on models, we end up with Lewis' semantics for counterfactuals. It has two special characteristics that makes it different from Stalnaker's original semantics for conditionals. First of all, if the field of $\leq_w$ is infinite, it allows for closer and closer A-worlds without a (set of) closest. Second, different possible worlds can be closest to w without being identical to each other, that is, ties are allowed. Stalnaker prohibits both possibilities by the *uniqueness assumption*: For every world w and non-empty proposition A, there is a unique closest world in A. This uniqueness assumption can be represented by the following two constraints on models:

(e) limit assumption : $A \neq \emptyset \Rightarrow \{w' \in A :\ \forall w'' \in A :\ w' \leq_w w''\} \neq \emptyset$

(f) trichotomy: $\forall w', w'' :\ w' <_w w''$ or $w'' <_w w'$ or $w' = w''$.[4]

Accepting the limit assumption [5] (or limiting our analysis to the finite case) and trichotomy, we can reformulate the semantics of counterfactuals in the following way: Let's call $M = \langle W, f\rangle$ a model structure in which W stands intuitively for the set of possible worlds and f for a function from worlds and propositions, subsets of W, to propositions, that satisfies the following conditions:[6,7]

[4]Note that by trichotomy, the equivalence classes induced by $\leq_w$ turn out to be singleton sets.

[5]I will always make the limit assumption.

[6]Or we define the selection function in terms of the similarity relation as follows:
$f_w(A) = \{w' \in A |\ \forall w'' \in A :\ w'' \leq_w w' \Rightarrow w'' = w'\}$.

[7]The principle behind this reformulation is simple, of course. Just say that $v \in f_w(\{v, u\})$ iff $v \leq_w u$.

(*a*) $f_w(A) \subseteq A$, success

(*b*) $f_w(A) = \{w\}$, if $w \in A$, strong centering

(*c*) if $f_w(A) \subseteq B$ and $f_w(B) \subseteq A$, then $f_w(A) = f_w(B)$, and

(*d*) $f_w(A)$ contains at most one member.

The proposition expressed by the conditional $A > B$ is the following set of possible worlds:

$$A > C \quad = \quad \{w \in W : \ f_w(A) \subseteq C\}$$

That is, $A > C$ is true in w iff C is true at every closest A-world to w, or A is impossible. Conditions (a)-(c) are assumed by both Stalnaker (1968) and Lewis (1973). The first condition guarantees that the truth value of a conditional is partly dependent on the truth value of the antecedent. Condition (b), the strong centering assumption, gives already some content to the notion of similarity. On the one hand it guarantees that if the antecedent is true, the conditional behaves like material implication, so modus ponens is valid. It assures that in case the antecedent is true, it is only the actual world that counts. On the other hand, it also validates the inference from $A \wedge C$ to $A > C$.[8] Condition (c) gives the most content to the similarity function. It says that the similarity function is independent of the antecedent (or conditional) to be evaluated. Lewis and Stalnaker disagree whether or not condition (d) should be given. The principle that corresponds with this condition, '$(A > B) \vee (A > \neg B)$', is known as the principle of the *conditional excluded middle* (CEM). It is proposed by Stalnaker to capture the intuition that we deny a conditional *if A then B* by *If A then not B*. Should we accept this principle? Stalnaker keeps saying *yes*, Lewis kept saying *no*. Lewis said *no*, because assuming (CEM) makes it unclear how to account for *might* counterfactuals, and anyway, ties are needed to account for Quine's sentences:

(107) a. If Bizet and Verdi were compatriots, Bizet would have been Italian.

b. If Bizet and Verdi were compatriots, Verdi would have been French.

How else can the fact that, intuitively, neither (107a) nor (107b) is true, be accounted for?

Stalnaker argued instead that *might* is epistemic and that it has normally wide scope over the conditional. Epistemic *might* doesn't say something about a single possibility, but about a whole information state, instead. So, if an information state **K** is represented by a set of possible worlds, K, and a selection function, f, a *might* counterfactual, $\Diamond(A > C)$ is acceptable in **K** iff there is a world in K in which $A > C$ is true with respect to the selection function. But what if there is only one world left? Stalnaker argues that even in that case he can account for *might* counterfactuals. The idea is that it might

[8] If we replaced strong centering by weak centering - $w \in A \Rightarrow w \in f_w(A)$ - this inference would no longer be valid.

be unclear what the right way is to select nearest possible worlds. We don't associate with a possible world a single set of spheres, but rather a set of sets, for each selection function a separate set. Conditional statements are not simply true or false in a world, but true or false in a world with respect to a selection function. Given that we have a set of selection functions, F, the notion of absolute truth and falsity is then defined in terms of supervaluation. A conditional is absolutely true (false) in a world, iff it is true (false) in this world with respect to all selection functions in F. This account makes it possible that some counterfactuals are neither true nor false in a world. Stalnaker suggests that this is what is going on with Quine's sentences.[9] In this way it also becomes possible to account for *might* counterfactuals: the counterfactual $\Diamond(A > C)$ is true in w if there is a selection function $f \in F$ such that $A > C$ is true in w with respect to f.[10]

Stalnaker (1980a) argued that even if there is in fact no penny in my pocket, although I do not know it since I did not look, the falsity of

(108) If I had looked, I *might* have found a penny.

can be explained by giving *might* wide scope. What should be done is that in this case we don't consider all the worlds compatible with my knowledge, but only those worlds compatible with my knowledge that I would have if I knew all the relevant facts. On this quasi-epistemic reading of *might* the conditional comes out false as wanted. It is important to note that even assuming this quasi-epistemic context of interpretation, this account still leaves open the possibility that the truth value of some conditionals can be indeterminate. Lewis (1973) argued, however, that the use of *might* in (108) has something to do with objective chance (indeterminism), which should not be modelled by epistemic uncertainty. But once one accepts the supervaluation account of van Fraassen (1974) and Stalnaker (1980a) for conditionals, one might argue that chance should be modelled by quantifying (put the probability measure) over (a set of relevant equivalence classes of) selection functions.[11]

Independent motivation for making the uniqueness assumption comes from the analysis of *only if* clauses (cf. von Fintel, 1994). To give an account of such clauses, we preferably analyze them compositionally in terms of the meaning of *only* and *if*, instead of treating *only if* as one connective. The main empirical constraint is that *B, only if A* seems

[9]For some additional arguments for Stalnaker's position, see Stalnaker (1984, ch. 7).

[10]Van Fraassen (1974) has proven that Lewis models which satisfy the limit assumption are equivalent to a family of Stalnaker models that satisfy the so-called *regularity* condition. The idea is that if $\{\langle W, f^s\rangle |\ f^s \in F\}$ is a family of Stalnaker models, then we can define a Lewis model $\langle W, f\rangle$, where for every $A \subseteq W$: $f_w(A) \stackrel{def}{=} \bigcup\{f^s_w(A) |\ f^s \in F\}$, if the family of Stalnaker models satisfies the regularity condition.

[11]This suggestion is appealing, especially because of the results of van Fraassen (1974) showing that, under certain conditions, Stalnaker models are Lewis models with hidden variables, and thereby making a connection between the hidden variable interpretation of quantum mechanics and Stalnaker's analysis of conditionals. We might for instance say that the chance of $A > C$ in w is n iff $P(\{\{f' \in F |\ f'_w(A) = f_w(A)\} |\ f \in F\}, \{\{f' \in F |\ f'_w(A) = f_w(A)\ \&\ f'_w(A) \subseteq C\} |\ f \in F\}) = n$, where F is the relevant set of selection functions.

to mean (also) the inverse of *If A, then B*: *if* $\neg A$*, then* $\neg B$. According to a well established tradition, if P is the constituent of A that is in focus *Only* A is true and appropriate iff A is true and A with P substituted for any of the relevant alternatives to P is not true. It is assumed that *only* A is interpreted with respect to an invisible contextual parameter C which contains a set of mutually inconsistent alternatives of which P is one, and that '$\text{only}_C(A)$' expresses the following proposition: $\{w \in A |\ \forall Q \in C :\ Q \neq P \rightarrow w \notin A[^P/_Q]\}$, where $A[^P/_Q]$ is A with P substituted by Q. The question now is how we should interpret *if A, then B* such that from the truth of *B, only if A* we can infer that *if* $\neg A$*, then* $\neg B$ is also true. First, note that in *B, only if A*, normally it is A that has focus. On the assumption that the set C should contain the relevant alternatives which are mutually inconsistent, this set should be $\{A, \neg A\}$. Suppose *if* means what the similarity approach predicts it to mean. In that case, the proposition expressed by *B, only if A* with respect to context C is true and appropriate in w iff $f_w(A) \subseteq B \ \&\ \forall D \in C[D \neq A \rightarrow f_w(D) \not\subseteq B]$. In our case this means that $f_w(A) \subseteq B$ and $f_w(\neg A) \not\subseteq B$. Note that in general we cannot infer from this proposition to $\neg A > \neg B$, because it might be the case that $f_w(\neg A) \not\subseteq B$, but $f_w(\neg A) \cap B \neq \emptyset$. Just consider the case where $W = \{u, v, w\}, A = \{w\}, B = \{v, w\}, f_w(\neg A) = \{u, v\}$. Once we make the uniqueness assumption, however, this will not happen.

5.3 The Ramsey test analysis

Lewis and Stalnaker analyzed conditional sentences semantically in terms of selection functions. They motivated the properties of those selection functions partly by purely empirical considerations. However, it would be nice if independent motivation for those properties could be given. This is tried in Stalnaker (1968, 1970a). In analytic philosophy it is traditionally assumed that objective modal concepts should be reduced to subjective ones. In the spirit of this tradition Stalnaker tried to understand the content of conditional propositions – and thus to motivate the properties of the selection functions – by making an analogy between the truth conditions of conditional sentences, and the way we evaluate conditionals with respect to our belief states. Stalnaker's (1968) analysis of conditionals is a generalization of a suggestion first made by Ramsey (1931) and therefore called the *Ramsey test* analysis of conditionals. Ramsey's pragmatic philosophy inspired him to reduce the meaning of sentences to beliefs. His natural suggestion was that the analysis of conditional sentences should be reduced to *conditional beliefs*. Stalnaker gives the following instructions for deciding whether you do or do not believe a conditional:

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is true. (Stalnaker, 1968, p. 102)

According to the above quotation, conditionals should be handled in terms of the relation between the actual belief state and the relevant hypothetical belief state. I will call this hypothetical belief state after the adding of A to belief state K and the required adjustments the state K *revised* by A.[12] If we represent a belief state by K, the Ramsey test analysis can be stated in the following way:

$A > C$ is true (accepted) in K iff C is true (accepted) in K revised by A.

Analyzing conditionals in terms of changing beliefs connects the interpretation of conditionals with that of inquiry.

> Inquiry is the process of changing [...] acceptance states, either by interaction with the world or by interaction between different acceptance states. Methodological policies are policies constraining such changes. To have a framework for describing methodological policy, we might assume that acceptance states have two components: a set of alternative possibilities representing the inquiry's current conception of the way the world is, and a change function representing his disposition to change what he accepts in response to new information. (Stalnaker, 1984, p. 99)

The properties of selection functions can be derived if it is assumed via the Ramsey test analysis that conditional propositions are projections of the methodological policies onto the world. Suppose that $\langle K, f\rangle$ is our belief state, where K is a set of alternative possibilities and f the change function. Now suppose that we learn so much about the world that K consists of one possible world only. In that case, f will coincide with the selection function used in the *semantic* analysis of conditionals. What properties can be derived from this methodology? That depends of course on how belief revision should be analyzed. But it is clear that for instance the condition '$f_w(A) = \{w\}$, if $w \in A$' can be motivated in this way: if you believe A already, you don't have to change your beliefs if you learn A. More generally, the Ramsey test analysis gives a motivation for why the notion of similarity should play a part for the analysis of conditionals. If you have to change your beliefs in response of new evidence, you want to give up only those beliefs for which the new evidence gives you reason to.

Let us assume with Stalnaker that there always exists a selection function f that assigns to every world w and proposition A a single possible world w', possibly the same as w. Suppose now that an information state is represented by $\langle K, f\rangle$, where K is the set of possible worlds in which all the accepted propositions are true, and f is the selection function. We can now define the change (revision) function of K, C, in the following straightforward way:

[12] Although this name is somewhat misleading, as Stalnaker reminded me: The hypothetical belief state that Stalnaker talks about in the above quotation need not necessarily be the belief state that would result from the agent's belief revision policies.

$$C_K(A) \quad = \quad \{f_w(A) : \ w \in K\}$$

If the selection function satisfies the constraints that Stalnaker (1968) argued for, it is justified to say that $C_K(A)$ is the minimal revision of K by A.

If, by the Ramsey test, belief in conditionals should be reduced to conditional belief, the analysis goes like this:

(a) $A > C$ is accepted in K iff $C_K(A) \subseteq C$

If the agent learns more and more about the world, ideally he once reaches an information state where K consists of only one possible world. In that case (a) comes down to (b):

(b) $A > C$ is true in w iff C is true in $f_w(A)$

The informal connection between belief in conditionals and conditional belief made in Stalnaker (1968) gives, arguably, a good motivation for those constraints on selection functions that Lewis and Stalnaker agree on. But it does not determine at all whether or not $f_w(A)$ should contain at most one member. Suppose we don't want to be committed to the assumption of trichotomy, then the definition of $C_K(A)$ changes into

$$C_K(A) \quad = \quad \bigcup\{f_w(A) : \ w \in K\}$$

In case K consists of only one world, (a) comes down to (c):

(c) $A > C$ is true in w iff C is true in all w' in $f_w(A)$

Above I defined the global change function, $C_K(A)$, in terms of selection functions on worlds. But that is not really what the Ramsey test analysis suggests, it should rather be the other way around. The selection function on individual possible worlds should be determined by how rational agents would change their global belief state by learning new information. Stalnaker (1968) tried to explain the proposition expressed by a conditional sentence as a projection of the epistemic notion of conditional belief onto the world. There is another tradition that tried to do something similar. According to the Bayesian account, objective modal notions like causality and chance should be reduced to their subjective counterparts. Stalnaker (1970a) intended to explain the properties of the selection function on individual possible worlds by making a connection between the Bayesian account of belief change and his 1968 analysis of conditionals. Moreover, by making this connection he wanted to settle the issue between Lewis and himself about the controversial Conditional Excluded Middle principle. To that we will turn now.

5.4 The Bayesian approach

Independently of the analysis of conditionals, another model of changing information states had been developed. In the epistemic *Bayesian tradition*,[13] information states (beliefs) of agents are modelled by probability functions and the (rational) change of information states is handled by *conditionalisation*.

Normally, conditional probability functions are defined in terms of singular ones by means of *conditionalisation*: $Pr^*_A(B) = Pr(B/A) = Pr(A \wedge B)/Pr(A)$, if $Pr(A) \neq 0$, and otherwise undefined. So in the Bayesian tradition too, the rational change of information states is defined in terms of *conditional beliefs*. Just like Stalnaker's account of conditionals was based on the analysis of minimal belief change, the analysis of minimal belief change in the Bayesian tradition also gave rise to an account of conditionals. Adams (1965, 1976) claimed that the assertability of an indicative conditional goes with its corresponding conditional probability. To base a logic of conditionals on probabilistic techniques, Adams based his analysis of conditionals on the following notion of ϵ-entailment:

> If S is a set of wffs, then A is ϵ- *entailed by* S iff for every $\epsilon > 0$ there is a $\delta > 0$ such that for all Pr, if $Pr(B) > 1 - \delta$ for each member B of S, then $Pr(A) > 1 - \epsilon$.[14]

Note that the resulting logic is non-monotonic. Learning new information can easily decrease certain (conditional) probabilities. In addition, principles like modus tollens and transitivity are invalid according to this notion of entailment. Adams claims that these principles should not be valid, not even for the analysis of indicative conditionals.

In the 20th century, probability theory has been widely used in the philosophy of science to construct logics of induction or confirmation. These logics have been developed to capture the inductive (causal) relations between propositions. As is well known, Popper argued that scientific theories are *not infallible*. Sometimes some propositions accepted before have to be given up because of some new phenomena. So, it cannot be that only tautologies are accepted.

If it is assumed that a belief state of a rational agent who accepts more than just logical tautologies is represented by a probability function, and that a proposition is accepted by this agent iff its probability function assigns to it probability 1, it follows that also some contingent propositions are assigned probability 1 by this probability function. In that case it makes sense to define the information state based on probability function Pr in the following way: $\Omega(Pr) = \{A : \ Pr(A) = 1\}$. Such an information state obeys the characteristic properties of what might be called an *acceptance state*. A set of propositions might be called an acceptance state iff (a) *conjunction closure* holds; if A and B belong to the set, $A \wedge B$ belongs to the set, too; (b) the set is *closed under* classical *entailments*, and (c) the propositions in the set are mutually *consistent*.

[13] The classical paper is Ramsey (1931), the best introduction is Jeffrey (1965).

[14] If S contains n propositions, you might think of δ as being ϵ/n.

However, according to classical Bayesians, a contingent proposition should never have probability 1. The problem with this view is that the conditions for being an acceptance state are no longer guaranteed if propositions are already believed if their subjective probability are higher than, for instance, 0.5.[15] Thus, if we say that $A \in \Omega(Pr)$ iff $Pr(A) > 0.5$. But on this assumption, we can easily find counterexamples to condition (a). Consider the following two statements about our next throw of an unbiased dice: *It will show 1, 2, 3 or 4*, and *It will show 3, 4, 5 or 6*. Their subjective probabilities are both higher than 0.5, but their conjunction certainly is not. Thus, if a sentence is accepted iff its probability is higher than 0.5, the set of accepted sentences cannot be closed under conjunction. But also condition (c) is not guaranteed. This is made clear by the (infinite) lottery paradox. Assume a lottery with (infinitely) many tickets. For each ticket i the probability that it will not be the winning ticket will be more than 0.5. But it is clear that this acceptance state is inconsistent, if it were closed under conjunction. To account for acceptance, or more in general for monotonic reasoning, Skyrms (1980a) does not use the notion of probability, but that of *resilience*. Resilience is a measure of *invariance under belief change*.[16] A proposition is accepted if its resilience is greater than 0.5. Skyrms (1980a, pp. 152-154) proves that in this way an acceptance state obeys the three conditions above.

The definedness condition on conditionalisation has the immediate consequence that if we want to represent beliefs in terms of standard probability functions, it becomes impossible to analyze *counterfactual beliefs*. The reason is that by this definition, Pr becomes a partial function: $Pr(A/B)$ is undefined when $Pr(B) = 0$. Unsatisfied with this partiality of Pr and motivated by his own philosophy of science, Popper (1959) gave conditions on probability functions, P, where P is always defined. This is done by taking conditional probability as basic. These functions are normally called *Popper functions*, but via Stalnaker (1970a) are also known as *extended probability functions* (epfs).

Popper argued that to account for the notion of *acceptance*, we should give up strict coherence. To capture the notion of acceptance, more propositions than only tautologies have to be assigned probability 1. By taking conditional probabilities as basic, Popper functions contain more information than standard probability functions. They contain the extra information of how one would change one's belief state by learning something that is inconsistent with one's present belief state. So, (one step) revision is built into the probability function. This extra information also captures invariance under belief change or *epistemic entrenchment*. Suppose two propositions are both believed, then one of the two can still be believed more strongly than another, if the latter would be given up earlier than the former. Why is one believed proposition given up earlier than another? The reason is that the former proposition is more strongly connected to other accepted propositions than the latter. Popper-functions contain information about the inductive and conceptual relations among propositions believed. If the correlation between the changing probability-

[15] Or any other real number in [0,1).

[16] Just like the notion of epistemic entrenchment that we will discuss later.

values of two propositions under different counterfactual assumptions is strong, it is likely that the events described by the propositions are connected with each other. The difference between logical tautologies and contingent propositions that are both accepted is then that the former will never be given up, while the latter will.

Harper (1976a) provided a justification for Popper functions. Note that the propositions believed can be recovered from epfs in the following way: $\Omega(P_A) = \{B : \ P(B/A) = 1\}$. Suppose now that every proposition is represented by the set of possible worlds in which it is true. It is then possible to compare the set of possible worlds that epf $P_\top$ assigns a non-zero probability, $K(P_\top) = \bigcap \Omega(P_\top)$, with the set of worlds that are consistent with P revised by A, $K(P_A) = \bigcap \Omega(P_A)$. Harper showed that the minimal revision modelled by Popper functions satisfies some intuitive conditions on minimal change of a certain belief state. These conditions gave rise to analyses of revision in purely qualitative frameworks.

Note that according to Popper, probability and entrenchment are two almost independent notions. This has motivated authors like Harper (1976a, 1976b), Spohn (1987), Gärdenfors (1988) and many more to model belief revision and epistemic entrenchment in a purely qualitative framework. In the most simple variants of these belief revision frameworks, an acceptance state is modelled by a set of possible worlds, K, and a belief revision function $*$. If we say that $\langle K, * \rangle$ is a belief state, and A any proposition, then K^*_A is called the revision of K by A, and this revision process is constrained by the following rules for minimal belief change:[17,18]

(R^*1) For any proposition A, $K^*_A \subseteq A$

(R^*2) If $A \neq \emptyset$, then $K^*_A \neq \emptyset$

(R^*3) If $K \cap A \neq \emptyset$, then $K^*_A = K \cap A$

(R^*4) If $K^*_A \cap B \neq \emptyset$, then $K^*_{A \wedge B} = (K^*_A) \cap B$

Stalnaker and Lewis used in their possible world analysis of conditionals a selection function defined on single possible worlds. Equivalently, their analysis of counterfactuals was based on an ordering relation between possible worlds; Lewis' system of spheres model, which has a unique world in the centre of the sphere. The selection function could then be defined in terms of this ordering relation. We have seen that in terms of these selection functions we could define a revision function on global belief states.

Revisions of conditional probability functions are based on an essentially different idea. The revision is primitively defined in terms of the *global* belief state, represented

[17]Note that the constraints say nothing about introspection. It is normally assumed that belief states are introspective, and thus it seems reasonable to assume that also a revised belief state should be introspective. The constraints (R^*1) – (R^*4) do not guarantee this, however. In fact, the new belief state will only be introspective if the proposition by which the old belief state was revised was already believed.

[18]The following 4 conditions are in possible world semantics equivalent to the 8 well known AGM postulates (see Gärdenfors, 1988).

by a conditional probability function. Let us call revision functions primitively defined in terms of global belief states *epistemic revision*. Just as the Lewis/Stalnaker analysis of conditionals could also be based on an ordering relation between possible worlds, Harper (1976b) showed that epistemic revision could be based on an ordering relation, $\preceq$, of possible worlds, too.[19] It is quite easy to see what condition this ordering relation has to satisfy to implement the same belief revision policy as $*$ does: $v \preceq w$ if $v \in K^*_{\{v,w\}}$. We can now check that this ordering relation is transitive and connected.[20] It can also be shown that if we take such an ordering relation $\preceq$ as primitive, we can define both K^*_A and K as follows: $K^*_A \stackrel{\text{def}}{=} \{w \in A |\ \forall v \in A : w \preceq v\}$ and $K \stackrel{\text{def}}{=} K^*_\top$, such that K^*_A satisfies the above mentioned constraints.

Remember that the logics of induction and confirmation were developed to capture the inductive (causal) relations between propositions. But that is exactly what the Stalnaker/Lewis logic of counterfactuals intended to capture, too. It is only natural that the following question sooner or later should arise: Are these ways of handling conditional beliefs, belief change, and inductive logics related, and if so, how? The answer to those questions would clarify something about how a set of possible worlds partly determines the selection function.

Stalnaker (1970a) made the following strong but also very natural proposal: the probability of truth of a conditional equals the conditional probability. This proposal does not only mean that conditionals in general could equally well be analyzed epistemically in the Bayesian tradition using epfs. But also that the two different analyses of revision (the *distributive* one of Stalnaker/Lewis and the *global* epistemic one) come down to the same. Assuming that the minimal revision of a belief state in terms of a similarity function is equal to the minimal revision of a belief state in terms of conditionalisation, his natural proposal was that $P(A > C) = P(C/A)$.

To be a bit more precise, let $\langle W, F, P\rangle$ be a probability space where W is a set of worlds, F a field of subsets of W closed under the Boolean operations, and P an arbitrary probability function closed under conditionalisation. Stalnaker implicitly assumed that '$>$' has a fixed interpretation. His hypothesis was that there is a binary connective '$>$' that behaves like a conditional such that for any probability space $\langle W, F, P\rangle$, such that for all $A, B \in F, A > B \in F$, and where the probability function can represent a rational agent's system of belief, it is the case that for all $A, B \in F : P(B/A) = P(A > B)$, if $P(A) > 0$ (Stalnaker does not really demand that $P(A) > 0$, but that is not important here). Because P is closed under conditionalisation, if P_C is the probability function that results from P by conditionalising on C, the hypothesis also says that $P_C(B/A) = P_C(A > B)$, if $P_C(A) > 0$. This proposal is known as *Stalnaker's hypothesis*.[21]

[19] See also Grove (1988).

[20] But it does not satisfy Lewis' strong centering condition. The reason is clear; the ordering relation does give rise to a system of spheres, but the centre of this sphere need not be a single possible world. See section 6.7 for more on this.

[21] See Hajek & Hall (1994) for a discussion of related hypotheses by Adams and others.

Stalnaker's main interest in this hypothesis was that, if this were true, it would give an independent argument in favour of his controversial conditional excluded middle (CEM) principle. Because for proposition A that has a non-zero probability, by definition $P(\neg B/A) = 1 - P(B/A)$. Assuming that Stalnaker's proposal were true, both $P(\neg(A > B))$ and $P(A > \neg B)$ would have the same value as $P(\neg B/A)$. From this we could then immediately derive Stalnaker's CEM.

Stalnaker's hypothesis can also be stated in a qualitative way. The hypothesis then says that there is a binary connective '$>$' that behaves like a conditional such that for any $K, A > B$ is accepted in K iff B is accepted in K^*_A. This in turn comes down to the hypothesis that there is a selection function f such that for any K and A, $K^*_A = \bigcup\{f_w(A)|\ w \in K\}$.[22]

5.5 Triviality

Stalnaker (1968) assumed that the correct account of conditionals should be based on the Ramsey test analysis: $A > B$ is accepted in K iff B is accepted in K revised by A. In Stalnaker (1970a) it is assumed that revision should be handled as in the epistemic Bayesian approach. These two assumptions together gave rise to the hypothesis, $P(A > B) = P(B/A)$. Lewis' triviality result showed, however, that this hypothesis is false for all but some trivial probability functions.[23] More in detail, Lewis showed that any probability function that satisfies Stalnaker's hypothesis and the constraint that the probability function is iterative:

$$(CSH) \quad P(A > B/C) \quad = \quad P(B/A \wedge C), \text{ if } P(A \wedge C) \neq 0$$

for any binary connective $>$, can only assign different probabilities to two different propositions. In Appendix D I will not only go over Lewis' proof, but I will also show how the extra constraint (CSH) follows from Stalnaker's hypothesis extended to conditional probabilities, and by the assumption that conditional probabilities satisfy the standard laws.

Over the years, a number of authors have strengthened and generalised Lewis' triviality proof for both probabilistic representations of information states and qualitative variants thereof.[24] The qualitative version of Stalnaker's hypothesis says that conditionals (i) state context independent propositions and (ii) should be handled by the Ramsey test

[22] In Gärdenfors (1988), $\langle \mathbf{K}, ^* \rangle$ is called a belief revision model, where $\mathbf{K}$ is a set of belief sets and * a revision function. Gärdenfors assumes that $A > B \in K$ iff $B \in K^*_A$ and that for any $K \in \mathbf{K}$ and any proposition A, the revision of K by A, K^*_A, is again an element of $\mathbf{K}$. In other words, what is assumed is that there is a $*$ such that for any $K \in \mathbf{K}$ and proposition A, the revision of K by A, K^*_A, is again an element of $\mathbf{K}$, and that $A > B \in K$ iff $B \in K^*_A$.

[23] For a clear exposition of Lewis' proof, and some much more telling results, see Hajek & Hall (1994).

[24] See Gärdenfors (1988) for a qualitative variant of the triviality proof, and Hajek & Hall (1994) for an overview. Gärdenfors' impossibility proof is a strengthening of Lewis' triviality proof, because there is no qualitative variant of the expansion by cases rule of probability functions.

analysis. The triviality proof (see Gärdenfors, 1988) is then based on the assumption that revision should be very much like conditionalisation in that it has to satisfy the following constraint:

$$(R^*4) \quad \text{If } K^*_A \cap B \neq \emptyset, \text{ then } K^*_{A \wedge B} = (K^*_A) \cap B.$$

Those two assumptions together lead to the qualitative version of (CSH) (where $K^*_A \models B$ means that B is accepted in K revised by A):

$$(CSH^*) \quad K^*_C \models A > B \text{ iff } (K^*_C \cap A) \subseteq B, \text{if } K^*_C \cap A \neq \emptyset$$

Note that (R^*3) is a special case of (R^*4). If a revision function satisfies (R^*3), the revision function is called *preservative*. What the triviality results at least show is that the following four conditions are not jointly satisfiable:

(a) conditionals should be analyzed via the Ramsey test,

(b) all conditionals state propositions,

(c) the conditional has a fixed interpretation,[25] and

(d) the revision function satisfies (CSH) or its qualitative variant (R^*4).

Note that the third condition is the basic assumption behind the original similarity account of counterfactuals, while the fourth condition is assumed in all global revision methods.

5.6 Reactions to triviality

Given that the four above principles are jointly responsible for the triviality results, it's clear that we can react in at least four ways to the results of Lewis and others. And indeed, this is what happened. Lewis (1975) showed that we could maintain the Ramsey test, and Stalnaker's hypothesis, if we give up (d), van Fraassen (1976) showed that Stalnaker's hypothesis could be preserved if we give up (c), Adams and Gibbard proposed to give up (b), while Lewis proposed to give up the Ramsey test analysis for conditionals in general. Stalnaker, gave up his own hypothesis, and also the hypothesis that conditionals have a fixed interpretation, but maintained the claim that conditionals should be given a uniform analysis. The present section gives a survey of these proposals.

[25] Meaning that the model structure contains only a single ordering relation, or selection function, in terms of which all conditional sentences have to be interpreted.

5.6.1 Imaging versus epistemic revision

After destroying Stalnaker's hypothesis, Lewis (1975) showed that we can keep the Ramsey test analysis for conditionals in general by giving up (CSH) and defining revision in terms of imaging. But what is imaging?

Imaging is a function of minimal belief change which uses not primarily the information available in the information state ordered by epistemic entrenchment (as in Stalnaker, 1970a, and Grove), but the similarity relation between *individual* possible worlds. The consequence is that it differs from conditionalisation (of normal or conditional probability-functions) in an interesting way. Here is the intuition behind it:

> Imaging P on A gives a minimal revision in this sense: unlike all other revisions of P to make A certain, it involves no gratuitous movement of probability from worlds to dissimilar worlds. Conditionalisation P on A gives a minimal revision in this different sense: unlike all other revisions of P to make A certain, it does not distort the profile of probability ratios, equalities, and inequalities among sentences. (Lewis, 1975)

Let us think of the probability functions as assigning probabilities to the (finitely many) worlds such that the probabilities add up to 1. Let us now assume (by the uniqueness assumption) that for every world w and proposition A there is a unique world $f_w(A)$. Given a probability function P and any possible A, there is a probability function P_A such that, for any world w':

$$P_A(w') \quad = \quad \sum_w P(w) \times \begin{cases} 1, \text{ if } f_w(A) = w', \\ 0 \quad \text{otherwise} \end{cases}$$

Lewis calls P_A the image of P on A, and says that P_A comes from P by imaging on A. Intuitively, the image on A of a probability function is formed by shifting the original probability of each world w over to $f_w(A)$. Then Lewis is able to prove (unsurprisingly) that $P(A > C) = P_A(C)$.[26]

What is interesting about imaging is that the preservation property for revision is no longer valid. Think of K as the information state before revision defined in the following way: $K = \bigcap\{B : \ P(B) = 1\}$, where each proposition B is represented by the set of possible worlds in which it is true. Revising the information state K by a proposition A that is

[26] Where $P_A(C) \quad = \quad \sum_w P(w) \times \begin{cases} 1, \text{ if } f_w(A) \in C, \\ 0 \quad \text{otherwise} \end{cases}$

Stalnaker didn't really show much interest in this last result. The reason should be obvious. Even though it also verifies the principle CEM, it can hardly be called independent motivation for it. And indeed, Gärdenfors (1982) showed that imaging can also be defined without the uniqueness assumption. The result is that the probability originally assigned to a world where A is not true is possibly spread over more than one world where A is true. This obviously reflects Lewis' analysis of counterfactuals, instead of Stalnaker's, in that it doesn't validate CEM anymore.

consistent with it does not result necessarily in an information state K' that is a subset of K. The reason is that even if $K \cap A \neq \emptyset$, the most similar A-world to a w in K that doesn't make A true, doesn't have to be an element of K. For this reason it might falsify a proposition that was verified by every element of K. Note that the state $K' = \bigcap\{B : \ P_A(B) = 1\}$ can also be determined by our earlier change function $C, K' = C_K(A)$. We can conclude that also the qualitative revision function C is not in general preservative, i.e. it does not generally satisfy (R^*4), the qualitative version of (CSH). Now we can understand why Stalnaker's (1970a) proposal was not justified; his assumption that the minimal revision of a belief state in terms of a context independent similarity function is equal to the minimal revision of a belief state in terms of conditionalisation was wrong, K^*_A need not be the same as $C_K(A)$.

Lewis (1975) showed that we can keep the Ramsey test analysis for counterfactuals, if we give up (CSH). Van Fraassen (1976) even showed that if we give up (CSH) a version of Stalnaker's hypothesis might still be true.

5.6.2 Van Fraassen

Discussing the triviality result, we have seen that any probability function that satisfies Stalnaker's hypothesis and principle (CSH), $P(A \Rightarrow B/C) = P(B/A \wedge C)$, if $P(A \wedge C) \neq 0$, for any binary connective $\Rightarrow$, will be trivial. In Appendix D one can see that one of the premises for deriving (CSH) was the assumption that $\Rightarrow$ has a fixed interpretation. Van Fraassen (1976) called this assumption *metaphysical realism* and proposed to give that up. Giving up the assumption that conditionals have a fixed interpretation, van Fraassen was able to prove that for every probability function there is a binary connective '$>$' such that it has the same meaning in both occurrences of the embedded conditional "$(A > B) > C$", and where both $P(A > B) = P(B/A)$, and CEM holds.[27] Note that van Fraassen's result is much weaker than that which first was proposed by Stalnaker. Stalnaker's (implicit) hypothesis was that there is a '$>$' such that for all $P, P(A > C) = P(C/A)$, whereas van Fraassen proved only that for every P there is a '$>$' such that $P(A > C) = P(C/A)$. Making '$>$' context dependent is compatible with van Fraassen's claim, but not with Stalnaker's hypothesis as originally intended. Van Fraassen's result is weaker than the original hypothesis in another respect, too. As shown in Stalnaker (1976a), the probability of $A > C$ cannot be equal to $P(C/A)$ such that '$>$' obeys Stalnaker's logic, even if '$>$' is made context dependent. Hajek & Hall (1994) showed that '$>$' cannot even obey Lewis' logic. Indeed, in van Fraassen's logic CE, the axiom that corresponds with the following constraint on selection functions is given up: if $f_w(A) \subseteq B$ and $f_w(B) \subseteq A$, then $f_w(A) = f_w(B)$, a constraint shared by Stalnaker and Lewis.[28]

[27] See also Gibbard (1980).

[28] For more discussion see Jeffrey & Stalnaker (1994), and for reasons to be suspicious, see Hajek & Hall (1994). It should be noted that Van Fraassen also had a second method of saving Stalnaker's hypothesis in the $\forall\exists$ form for triviality, viz. by restricting the hypothesis to a limited class of conditionals. In that

5.6.3 Two kinds of belief change

If conditionalisation and expansion are special kinds of revision, we might say that van Fraassen proposed to give up the assumption that revision should obey (R^*4), or its corresponding probabilistic version. But this seems to me a very unnatural reaction, because this constraint seems to capture exactly what is going on if we change our beliefs by learning new information. Moreover, the principle also seems to be needed to account for (embedded) indicative conditionals. Giving up preservativity or the assumption from which it follows doesn't seem to be the right way to go. Indicative conditionals are only appropriately asserted in a given context if their antecedents are consistent with the context, if a context is represented by a set of worlds. To interpret the consequent, we should consider only worlds in the context

Contrary to van Fraassen, Stalnaker responded to the triviality result of Lewis by giving up his hypothesis and by arguing that, on second thought, the probability of truth of a counterfactual *should* not be equal to its corresponding conditional probability. Remember that the Bayesian account of probability is purely epistemic in nature. So $P(C/A) > P(C)$ means that A is evidentially relevant for the acceptance of C. But if his original analysis of counterfactuals is an appropriate analysis of causal relations and if Stalnaker's proposal were true, evidential relevance would be equal to *causal relevance*. But this is clearly not true and some puzzles in Jeffrey's (1965) purely evidential decision theory made this clear. According to Jeffrey's decision theory, actions are evaluated according to the probability the deliberator assigns to the desired state conditional on the proposition expressed by the action. The conditional probability $P(C/A)$ models the evidential relation the agent sees between A and C; if $P(C/A)$ is high, the agent would assign a high probability to C, if he would learn the news that A is the case. Obviously, if A causes C, $P(C/A)$ would be high, but the problem is that $P(C/A)$ might also be high in cases where A does not cause C, but where both are caused by a common cause. Stalnaker (1980b) gave the following example: Suppose that the correlation between smoking and lung cancer was not due to the consequences of smoking through the lungs, but due to a common genetic factor that causes both the tendency to smoke and the tendency to develop lung cancer. In that case there is no reason for agents to stop smoking in order to prevent lung cancer, although the probability of getting a lung cancer conditional on smoking is high. Stalnaker concluded that causal relevance, the kind of relevance needed to evaluate one's actions in a deliberation, should not be modelled by conditional probabilities of consequences with respect to actions. He suggested that, instead, the use of conditional probabilities in Jeffrey's theory should be replaced by the probabilities of their counterfactuals expressed. This suggestion has been worked out by various authors and resulted in *causal decision theory* (see Gibbard & Harper (1978)).

With the distinction between evidential and causal decision theory, there corresponds

---

case, the logic for conditionals is allowed to be as strong as Stalnaker's logic.

a distinction between two ways of changing one's belief state. Conditionalisation, or preservative revision, is supposed to mirror the way a rational agent would change his belief state if he would learn new information, while imaging is supposed to mirror the way a rational agent would change his belief state if he, or somebody else, would do a certain action.[29]

Stalnaker argued that $P(A > C)$ and $P(C/A)$ should in general not be the same. Global revision, and distributive revision by imaging reflect a different intuition. A number of authors have suggested that this difference corresponds to a difference between indicative conditionals and counterfactuals. According to this suggestion, an assertion of an indicative conditional mirrors the conditional probability the speaker assigns to the consequent with respect to the antecedent, while what is expressed by a counterfactual is less directly dependent on the speaker's current belief state. In the following sections we will discuss various ways in which these suggestions have been implemented.

5.6.4 Adams

According to Adams (1970) there exists a difference between indicative and subjunctive conditionals. He motivated this distinction by noting that if *Oswald* is in focus there is a difference between accepting (109a) and (109b):

(109) a. If Oswald didn't shoot Kennedy then someone else did.

b. If Oswald hadn't shot Kennedy, someone else would have.

If we learn that Oswald did not shoot Kennedy, we would immediately accept that somebody else did, but it is not so clear that we would accept that someone else would have killed Kennedy if Oswald hadn't shot him. Adams proposed already in the sixties that $P(B/A)$ should not be equated with the probability of *truth* of $A > B$, but rather with its *assertability*. This is not only the case for indicatives, but also for counterfactual conditionals. The difference between (109a) and (109b) is then explained by choosing a different probability function[30] for the analysis of an indicative conditional and its corresponding counterfactual. Because $P(B/A)$ is not equated with the probability of truth of $A > B$, Adams need not assume that $A > B$ states a proposition. It can be argued that what the triviality result really shows is that conditionals do not express propositions. The problem with this suggestion is that if conditionals no longer express propositions, it is not clear anymore how to account for *embedded* conditionals.

[29] In Katsuno & Mendelzon (1991) the qualitative version of conditionalisation is called the *revision* of a belief state, and by the *update* of a belief state is meant the qualitative version of imaging.

[30] He proposed in Adams (1976) that for counterfactuals, not the current, but a *prior* probability function is relevant. For a somewhat different proposal, see Skyrms (1994).

5.6.5 Lewis

Also for Lewis (1975) the distinction between indicative and subjunctive conditionals is a real one. He was happy to give up (in fact, never defended) the assumption that we should analyze all conditionals in a uniform way by the Ramsey test analysis.[31] Lewis never accepted the global revision approach for subjunctive conditionals. For his way of handling counterfactuals, and the probability thereof, the triviality result was not disturbing. On the contrary, the triviality proof showed that the independent motivation for principle CEM, that he had argued against before, was no good. Lewis also did not agree with Stalnaker that indicative conditionals should be handled in the same way as counterfactuals. According to Lewis the difference between the two corresponds to a *semantic* distinction. But he did not agree with Adams that indicative conditionals do not express propositions. Indicative conditionals state propositions, but should be analyzed in terms of material implication. To account for the paradoxes of the material implication he proposed to rely on Grice. He was prepared to concede to Adams that the assertability of indicative conditionals goes by conditional probability.[32] But claiming that the truth conditional content should be handled by the material implication enables him to account for iterated conditionals.

Lewis' analysis of indicative conditionals is, however, not very natural, because it treats the antecedent and the consequent of indicative conditionals *symmetrically* with respect to *truth*, but *asymmetrical* with respect to *assertability*. But this is problematic if there are examples where a conditional has intuitively a different assertability value to another sentence that by the material implication account of conditionals is truth-functionally equivalent with it. Examples of this kind have been given by various authors, but the proponent of the material implication account can always argue that the difference is not due to truth-conditional content, but to the *form* in which this content is asserted. But even this defense strategy doesn't work once we embed two such clauses into a larger sentence that intuitively has a different assertability or even truth value. Gibbard (1980) has given such an example, but maybe the greatest difficulty for a Gricean account for indicative conditionals is given by Grice (1989) himself. Consider the case where Yog and Zog play chess, Yog has white 9 out of 10 times, and draws are not allowed. We don't know who won what game, but we do know that of the hundred games they played up to now, Yog won 80 times when he had white and lost all 10 times when he had black. Intuitively, the following two assertions are true of any one of the hundred arbitrary games they played:

[31] Also Gärdenfors (1988) responded to the triviality result by giving up the assumption that all conditionals should be analyzed via the Ramsey test.

[32] Note that if indicative conditionals are analyzed by material implication, probability of truth and conditional probability equal each other only in extreme cases. On the other hand, Lewis (1975) showed that the conditional probability equals the probability of the material implication minus the probability of those cases in which asserting the conditional would be misleading: $P(B/A) = P(A \rightarrow B) - [P(\neg A) \times (P(A \wedge \neg B)/P(A))]$.

(110) a. If Yog had white, there is a probability of 8/9 that he won.

b. If Yog didn't win, there is a probability of 1/2 that he didn't have white.

The problem for the material implication account is that it cannot account for the truth of these assertions. If the probability operator has scope only over the consequent, we again have the well known paradox of material implication. If the probability operator takes scope over the whole conditional, the material implication account would predict that the embedded sentences of (110a) and (110b) are truth-functionally equivalent. But how can that be if their assigned probability is different?[33] To treat both assertability and truth in a similar way, it appears natural to use Belnap's three-valued (and two-dimensional) analysis of conditionals to determine the truth value of indicative conditionals. In this way, for most indicative conditionals at least, conditional probability equals its probability of truth (see Skyrms, 1980a, p. 89).

5.6.6 The preservativity principle

Contrary to Lewis, Stalnaker claimed that the similarity analysis of conditionals can be used for both counterfactuals and indicative conditionals. However, this gives rise to a problem; the following argument is not valid:

(111) a. Either the butler or the gardener did it.

b. Therefore, if the butler didn't do it, the gardener did.[34]

To account for this intuitively valid argument, Stalnaker introduced the notion of *reasonable inference*, a pragmatic relation between speech acts instead of the semantic relation of entailment between propositions. C is a reasonable inference of $A_1, ..., A_n$ iff the content of C is entailed by the context resulting from the initial context updated by $A_1, ..., A_n$, provided that for each $i \leq n$, the assertion A_i is made in an appropriate initial context.[35] The above direct inference is a reasonable inference if the following assumptions are made for being appropriate contexts for disjunctions and indicative conditionals:[36]

[33]The problem is, of course, that contraposition is valid according to the material implication account. There are other examples suggesting that contraposition should not be valid for indicative conditionals: from *If it is after 3 o'clock, it is not much after 3 o'clock* we don't infer *If it is much after 3 o'clock, it is not after 3 o'clock* (Nute, 1984, p. 428).

[34]If the conditional is analyzed as the material conditional, the argument is predicted to be valid. But Stalnaker rejects this analysis for indicative conditionals because it leads to a lot of other well known problems.

[35]See the appendix to Stalnaker (1975) for more details.

[36]If all indicative conditionals obey (a), it gives rise to the following appropriateness condition: It is appropriate to make an indicative conditional statement or supposition only in a context which is compatible with the antecedent. For a motivation for this principle, see Stalnaker (1975). Two other inferences that are invalid according to Stalnaker's semantics for conditionals, contraposition and the hypothetical syllogism, turn out to be reasonable for indicative conditionals.

(a) If an indicative conditional is being evaluated at a world in the context set, then the world selected must, if possible, be within the context set as well.

(b) A disjunctive statement is appropriately made only in a context which allows either disjunct to be true without the other.

Suppose that $A \vee C$ is appropriate in a given context. It follows that $C \wedge \neg A$ is compatible with the context set that represents the presupposed information. If then $A \vee C$ is added to the context set, the antecedent of the conditional *if* $\neg A$*, then* C will be compatible with the new context set. Because all $\neg A$-worlds in the context set are C-worlds, and because by (b) the selected $\neg A$ world will be a world in the context set, the inference in (111b) is reasonable.

Let K be a presupposition state, then we might say that for indicative conditionals with antecedent A, Stalnaker's appropriateness condition (a) has the following principle as consequence:

> A-worlds in K are to be selected as nearer to worlds in K than any A-world outside of K.

This principle seems to be the only reasonable assumption to make for an appropriate analysis of indicative conditionals. The principle is known as the principle of *preservativity*. We can follow Harper (1976b) and Morreau (1992) and implement this principle by relativising the selection function to the belief state. Given the definition of $f_w(A)$, we can relativise our selection function to a context K in the following way:

$$\begin{aligned} f_w^K(A) &= f_w(A \cap K), \text{ if } w \in K \text{ and } A \cap K \neq \emptyset \\ &= f_w(A), \text{ otherwise.}^{37} \end{aligned}$$

Now we can determine what proposition is expressed by the indicative conditional *If* A*, then* B in context $K, A >_K B$:

$$A >_K B \quad = \quad \{w \in W : \ f_w^K(A) \subseteq B\}.$$

In terms of the relativised selection function, we can also define the following context dependent revision function, $C'_K(A)$, the revision from K with A:

$$C'_K(A) \quad = \quad \{f_w^K(A) : \ w \in K\}$$

As especially made clear in Morreau (1992), if we accept the preservativity principle for selection functions, when the belief state changes, the selection function changes too. Let $\mathbf{K}$ be a belief state represented by $\langle K, f \rangle$, where K is a set of worlds and f a selection function.

[37] The ordering relation in terms of which the selection function is defined still obeys centering, transitivity, and connectedness.

When we revise $\mathbf{K}$ by A, the new belief state will be of the following kind: $\langle \bigcup\{f_w^K(A) : w \in K\}, g\rangle$, where $g = f^{\bigcup\{f_w^K(A):\ w\in K\}}$, a function from worlds and propositions to propositions, that satisfies not only success and weak centering, but also the preservativity principle with respect to $\bigcup\{f_w^K(A) : w \in K\}$. Morreau (1992) showed that if the preservativity principle is assumed, this extra dynamic element of belief change can handle examples (Hansson's example, and Tichy's example) that are problematic if it is assumed that the selection function doesn't change.

5.6.7 Gibbard

A number of authors have observed that the interpretation or assertability of an indicative conditional is much more context dependent than that of a subjunctive one. To account for this difference, Gibbard (1980) argued that while subjunctive conditionals can express context independent propositions and should be analyzed in terms of Lewis' and Stalnaker's original similarity account, indicative conditionals are more closely related to the epistemic state of the agents who utter them, and should be analyzed via the Ramsey test analysis. The latter suggestion can be implemented in two ways. Either we follow Adams and Belnap and claim that by uttering indicative conditionals we do not always express propositions, but instead make conditional assertions. What is asserted then depends on what is believed by the speaker. The other possibility is that we still demand that indicative conditionals always express propositions, that those conditionals are handled via the Ramsey test, but that we give up the assumption that the conditional has a fixed interpretation. We have seen that Stalnaker suggested something like the latter approach. Gibbard argues that the first approach is to be preferred, because – contrary to Stalnaker's analysis – it can account for the paradoxical fact that people who believe the conditional *if A, then B* can come to accept the opposite conditional *if A, then not B* and learn something from it, without having to revise their old belief state.

One of the central features of Stalnaker's (and Lewis') conditional logic is the *principle of conditional non-contradiction*, the assumption that $A > B$ is inconsistent with $A > \neg B$. This in distinction with the material implication: out of $A \rightarrow B$ and $A \rightarrow \neg B$, you cannot derive a contradiction but instead conclude $\neg A$. Gibbard (1980) has given a very nasty example that shows a problematic aspect of the principle of conditional non-contradiction:

> Sly Pete and Mr. Stone are playing poker on a Mississippi riverboat. It is now up to Pete to call or fold. My henchman Zack sees Stone's hand, which is quite good, and signals its content to Pete. My henchman Jack sees both hands, and sees that Pete's hand is rather low, so that Stone's is the winning hand. At this point the room is cleared. A few minutes later Zack slips me a note which says "if Pete called, he won", and Jack slips me a note which says "if Pete called, he lost". I know that these notes both come from my henchmen, but do not know which of them sent which note. I conclude that Pete folded. (Gibbard, 1980)

Gibbard argues that if both utterances express propositions, both of them should be accepted as true. But this is inconsistent with the principle of conditional non-contradiction. Because Gibbard does not believe that conditionals should be handled as material implication, he concludes that indicative conditionals do not express propositions. Instead, they are conditional assertions that mimic the probability the speaker assigns to the consequent conditional on the antecedent. He also argues that this non-propositional account of indicative conditionals has an extra advantage: it explains why many embeddings of indicative conditionals don't seem to make sense. Embeddings to the right, $A > (B > C)$ are not so problematic for the probabilistic account, if it is assumed that they are equivalent to $(A \wedge B) > C$.

5.6.8 A unified account

Assuming that indicative conditionals do not state propositions is problematic, however. Indicative conditionals embedded to the left are then difficult to handle, although (at least sometimes) they do make sense.

(112) If the cup broke if dropped, then it was fragile. (Gibbard, 1980)

Moreover, it doesn't seem very plausible to assume that indicative conditionals should be analyzed so differently from subjunctive conditionals. It is unwanted because it cannot be explained anymore why both kind of conditionals use the same words, combine with the same functions (even if, only if , if..might) in similar ways, and can be paraphrased in the same way (cf. Stalnaker, 1984). Note that the probabilistic account is a global account towards conditionals. It follows that conditionals are accepted or not with respect to a whole belief state. But that would mean that we can never learn anything from accepting a conditional. To account for the latter, it seems we have to assume that a conditional expresses a proposition. Most important, however, is that if we agreed with Gibbard, it would be much harder to explain, following the pragmatic tradition, the (seemingly) objective concept of counterfactuality in terms of the epistemic notion of conditional belief.

It seems that in order to give a unified account we have to follow Stalnaker and use conditional logic for both subjunctive and indicative conditionals. But the threat of Gibbard's problem remains.

If we want to analyze all conditional sentences as propositions, have a uniform Ramsey test analysis of the conditional, and demand that revision should satisfy preservativity, we have to give up the assumption that the conditional has a fixed interpretation. The most straightforward and clarifying way in which this can be done is to follow Harper (1976b).

5.7 Harper's principle and iterated revision

To account for iterated revision by learning new information, we would like our change function to obey all of $(R^*1)-(R^*4)$. One of the nice things about the original Lewis/Stalnaker account is that nested conditionals, or iterated revisions, do not give rise to interpretation problems. A similarity relation is given once and for all. A conditional, like any other sentence simply denotes a proposition. However, we have seen that revision by imaging does not guarantee that (R^*4) or its probabilistic variant will be obeyed. Harper (1976a) tried to construct non-trivial models of iterated belief change by restricting Stalnaker's hypothesis to the level of certainty: $P(A > B) = 1$ iff $P(B/A) = 1$. Stalnaker (1976b) showed, however, that by making the assumption that '$>$' has a fixed interpretation, and by accepting the following limited version of (CSH): if $P(A \wedge C) \neq 0 \Rightarrow P(A > B)/C) = 1$ iff $P(B/A \wedge C)) = 1$, that is, by accepting (R^*4), the conditional connective collapses into material implication.[38]

As we have concluded earlier, if we want to analyze conditionals via the Ramsey test paradigm, we have to make the interpretation of the conditional dependent on the particular acceptance states. In section 5.6.6 we have already seen how that can be done: make sure that the selection function obeys the preservativity condition. However, to account for iterated revision, or nested indicative conditionals, this won't quite do. Even by accepting the preservativity condition it is still not guaranteed that the following more general constraint is met:

$$(R^*4) \quad C'_K(A) \cap B \neq \emptyset \text{ only if } C'_K(A \wedge B) = C'_K(A) \cap B.^{39}$$

But it is this constraint that is needed to handle iterative revision, and thus indicative conditionals nested to the right.

Van Fraassen (1976) was able to make the meaning of the conditional context dependent, and in principle still could account for embedded conditionals. However, he had to give up (R^*4) too. Thus, the question arises whether it is possible to account for iterated revision, and thus for embedded conditionals, without giving up (R^*4)?

Harper (1976b) proved that we can, without the consequence that '$>$' is material implication. The price he had to pay, however, was that conditionals are even more context dependent than in van Fraassen's construction. In van Fraassen's theory, both of the connectives in a conditional like $A > (B > C)$ have the same meaning, while for Harper the two connectives have a different meaning. But exactly this made it possible to obey (R^*4). The way he built this context dependence into the meaning of the connective, into the selection function, is to make the Lewis/Stalnaker notion of similarity dependent on

[38] See also Gibbard (1980) for a closely related result.

[39] Consider the three logically independent propositions P, Q and R, and the eight worlds representing their possible combinations. Then we consider the following three propositions, $K = Q \cap R, A = \neg Q \cap (\neg R \cup P)$ and $B = \neg Q \cap (\neg R \cup \neg P)$. It can now be checked that $C'_K(A) = \neg Q \cap (P \equiv R), C'_K(A) \cap B = \neg Q \cap P \cap R$, but $C'_K(A \wedge B) = \neg Q \cap \neg R$.

the information state. This dependence is made so systematic that iterated revision is not problematic anymore. Harper makes the meaning of the conditional context dependent by accepting the following principle, that I will refer to as Harper's principle (HP):

> (HP) Only propositions decided by K should count in determining comparative similarity relative to K.

Harper defends this principle as follows:

> If one reflects on the role of Ramsey test conditionals the new principle is very plausible. As an acceptance context the total content of K is given by the propositions it decides, therefore it is just these propositions that should form the basis of judgment of comparative similarity relative to K. (Harper, 1976b, p. 130)

To formalize the principle, first a definition. For subsets S of W, belief states K, and worlds x and u, let $S^x_u K$ be the set of K-decided propositions in S on which x and u differ:

$$S^x_u K \;=\; \{A \in S :\; (K \subseteq A \text{ or } K \subseteq \neg A) \text{ and } ((x \in A \;\&\; u \notin A) \text{ or } (x \notin A \;\&\; u \in A))\}$$

Harper formalizes Harper's principle, by adopting the following constraint on $\preceq^w_K$, the relativised comparative similarity relation:

$$(HP) \quad \text{If } S^x_u K = S^y_u K \text{ and } S^x_v K = S^y_v K, \text{ then } u \preceq^x_K v \text{ only if } u \preceq^y_K v$$

If u and v both differ from x on exactly the same K-decided propositions in S on which they differ from y, then their comparative similarity to x relative to K must agree with their comparative similarity to y relative to K.

Now we define a relational measure of nearness based on the assumption that only the propositions that are decided by K determine similarity. How is that done? We can say that u is at least as similar to w as v iff the cardinality of the K-decided designated propositions on which u differs from w is less than or equal to the cardinality of the K-decided designated propositions on which v differs from w. In the simplest way we can take this set of designated propositions to be the set of atomic propositions, but you might also take this set to be any other set of arbitrary subsets of W.[40] More formally, we can define relative nearness in the following way: $u \preceq_K v$ iff $|S^w_u K| \leq |S^w_v K|$, for any $w \in K$, where S is the set of designated propositions that potentially determine similarity. From this definition it follows that the similarity relation obeys weak centering,[41] transitivity,

[40] See Harper (1976b) for details. For instance, we can order the elements of S, first we look only at elements of S, that correspond with lawlike sentences, and if that does not discriminate enough, we can also look at other propositions.

[41] Because only the K-decided propositions count in determining similarity, for any world w in K, $f^K_w(\top) = K$. Strong centering cannot be assumed anymore. Thus, we can no longer infer $A > C$ from $A \wedge C$. According to Adams (1976) we shouldn't; with it we cannot account for *explanatory uses* of counterfactuals.

connectedness, the limit assumption, and that Harper's principle is true. This similarity relation gives rise to a selection function and a system of spheres. The selection function is defined as follows:

$$f_w^K(A) \quad = \quad \{v \in A |\ \forall u \in A :\ v \preceq_K u\}$$

This selection function does not satisfy strong, but weak centering, $w \in A \Rightarrow w \in f_w^K(A)$, if $w \in K$. Now we can account for iterated revision, because if a set S of propositions that potentially determine similarity is assumed, from any set of possible worlds we can determine a system of spheres that belongs to it. Thus, we might say, qualitative revision is no longer a function from a system of spheres to a set of possible worlds, as in Grove (1988), but a function from a system of spheres to a system of spheres, or a function from an ordering, or similarity, relation to another ordering relation.[42]

We know already that if the similarity relation obeys transitivity, connectedness and weak centering, the system of spheres model will look similar to Grove's model which he used to analyze belief revision. Indeed, let the change function determined by accepting Harper's principle be denoted by C''. Then it can be proved that $f_x^K(A) = f_y^K(A) = C''_K(A)$, if $x, y \in K$, and that $C''_K(A)$ satisfies $(R_1^*) - (R_4^*)$. And this gives us exactly what we wanted for indicative conditionals. C'' is preservative, and the assumption that the similarity function is context independent is given up. In this way indicative conditionals are made heavily context dependent, without giving up the assumption that they express propositions and should be handled by the Ramsey test analysis.[43]

Note that the original Lewis/Stalnaker notion of similarity is a special case of Harper's construction. For Lewis and Stalnaker, the set K that represents the belief state is simply a singleton set. Thus, given a set $S, u \leq_w v$ iff $|S_u^w\{w\}| \leq |S_v^w\{w\}|$. Because a world decides all propositions, all propositions in S actually determine similarity. Lewis and Stalnaker always argued that the notion of similarity is context dependent. In our terms we might say that Stalnaker (1968) and Lewis (1973) already made the selection function dependent on what propositions potentially determine similarity, and that Harper showed that this selection function can also systematically depend on what is believed. Where Lewis and Stalnaker could already account for the fact that two agents whose beliefs are compatible with each other could justifiably assert two incompatible conditionals, because they assumed different ways of selecting closest worlds, Harper can also explain such cases by pointing to the difference of information available to the two agents.

[42]For a different account of iterated revision, see Spohn (1987).

[43]Let K be $\bigcap \Omega(P)$ for a particular Popper function P. Very simplistically, we can then define for all non empty K and A : $P(B/A) = |B \cap C''_K(A)|/|C''_K(A)|$. Suppose now simplistically that $P(A) = |C''_K(A)|/|C''_K(\top)|$, then a relativised and weaker version of Stalnaker's hypothesis: $P(A >_K B) = 1$ iff $P(B/A) = 1$, is true for all Popper functions P. Harper's result does not depend on the particular way we defined probability. That the stronger result, $P((A >_K B)/\top) = P(B/A)$, cannot be proved if the logic of $A > B$ is Stalnaker's logic $C2$, is proved by Stalnaker (1976a), where he proves that Lewis' triviality result for Stalnaker's logic does not depend on the assumption that conditionals have a context independent fixed meaning.

5.8 Gibbard's problem revisited

Let's now go back to Gibbard's poker game example. To account for the conclusion in the poker game case that Pete folded is not so difficult. Let A, B, C, D, E and F be the following propositions:

A: Pete called,
B: Pete won,
C: Stone's hand is quite good,
D: Pete knows Stone's hand as well as his own,
E: Pete is disposed to fold on knowing that he had the losing hand, and
F: Pete had the losing hand.

Let I be our belief state and let $K(z, w)$ and $K(j, w)$ be the belief states of Zack and Jack respectively in w. We know that Zack believes C, D and E, so $\forall w \in I : K(z, w) \subseteq C \wedge D \wedge E$. It follows that for all w in $I, C'_{K(z,w)}(A) \subseteq B$. We also know that Jack believes C and F, so $\forall w \in I : K(j, w) \subseteq C \wedge F$. It follows that for all w in $I, C'_{K(j,w)}(A) \subset \neg B$.

But how can we conclude from both assertions *If Pete called, he lost* and *If Pete called, he won*, that Pete folded without knowing who made what statement? We assume that both are justified in claiming what they did, because the premises on which they base their conclusion are true and (by Gricean reasoning) they believe that what they say is true. I know D, E and that Jack knows both hands. I argue as follows: Suppose Pete had the losing hand, by D he knows he has the losing hand, and by E he folded. Now suppose Pete had the winning hand, by D he knows he has the winning hand, so the conditional $A > B$ would be true. Either Jack or Zack gave me a note which said $A > \neg B$. By looking at the context dependence of $A > \neg B$, we can not only determine what is expressed by the sentence once we know who wrote the letter, but once we know enough about the belief states of the possible writers and we assume some reasonable principles of communication, we might also be able to determine in what context we were, that is, determine who wrote the letter. It is clear that if Pete has the winning hand, the writer of $A > \neg B$ could not be Jack. The reason is that just like Pete, also Jack knows both hands. But I can also infer that Zack could not have written the note $A > \neg B$, if Pete had the winning hand. The reason is that I know that Zack knows that Pete knows both hands, and that both Zack and I assume E. So, the fact that either Jack or Zack wrote the note $A > \neg B$ is incompatible with the assumption that Pete had the winning had. So Pete must have had the losing hand, and so, by E, he folded.

So, also without accepting the material implication account of indicative conditionals we can infer that Pete folded. But this was not the main threat of Gibbard's example. His example was meant to show that it makes no sense to claim that the respective conditionals express propositions, and that even if we don't know so much about the belief states of Jack and Zack we still can infer that Pete folded if we assume that the two messages are reliable.

We have seen that Stalnaker (1975) made the selection function, and thus conditionals, context dependent. But that doesn't help as long as the meaning of the conditionals depends on the same context. What is needed is that the proposition expressed by the conditional depends on the beliefs of the *speaker*. Because the speakers can have different beliefs, the meaning of the same conditional sentence can still be different. But the problem is that sometimes we don't know who the speaker is, so – according to Gibbard – there can be no proposition expressed by an indicative conditional. But as Gibbard notes, the same thing can be true for sentences with indexicals. The proposition expressed by the sentence written on a postcard sent without an addressee with the message *I'm doing fine*, depends in the same curious way on who the unknown sender of the postcard is.

This suggests that Gibbard's problem should be solved in the same way as an utterance which uses referential expressions, but for which it is not clear what the actual referent is: *diagonalisation*.

We have seen earlier that the proposition expressed by the conditional $A > C$ with respect to context K is $\{w \in W |\ f_w^K(A) \subseteq C\}$. But we have implicitly assumed that context K represents what is *presupposed* in the *actual* world. Gibbard's example suggests that for the analysis of indicative conditionals we should not look at the presupposition state, but rather at the *speaker's belief state*, while diagonalisation suggests that the relevant information state should not (only) be the one in the actual world. Combining both we can say that if we know that a uttered $A > C$, the diagonal proposition expressed is $\{w \in W |\ f_w^{K(a,w)}(A) \subseteq C\}$. Unfortunately, in Gibbard's example we don't know who uttered what statement, nor do we know who believed what. As a result, we have to distinguish more cases if we want to make use of diagonalisation.

Let us first sketch the situation. In the initial situation for me, for Jack, and for Zack, there are three possibilities; Pete folded, Pete called and won, and Pete called and lost. Let's call those three situations w_1, w_2 and w_3 respectively. As far as we know, if Pete calls he might either win or loose. After Zack and Jack looked into the cards, their information states changed. Assuming that the utterers of *If Pete called, he won* and *If Pete called, he lost* were justified in claiming what they did, their belief states can be represented in all the worlds above by $\langle\{w_1, w_2\}, f\rangle$ and $\langle\{w_1, w_3\}, g\rangle$, respectively, where the two selection functions obey the preservation principle with respect to the belief state to which they belong, viz.: $f_{w_1}(\text{Pete called}) = \{w_2\}$, and $g_{w_1}(\text{Pete called}) = \{w_3\}$.

To account for diagonalisation we need to make a distinction between the different roles of context and index (world). The context determines *what* is expressed by a sentence, while the index determines whether what is said is true or not. If someone writes me a note which says *If A, then B*, it depends on the context what proposition is expressed by it. Two kinds of contexts are relevant in our example, one in which the belief state of the utterer of the conditional can be represented by $\langle\{w_1, w_2\}, f\rangle$, and one in which the belief state of the utterer of the conditional can be represented by $\langle\{w_1, w_3\}, g\rangle$. Thus, different contexts correspond with different selection functions; f and g. If we don't know in what

context we are, we don't know who slipped the note, but still want to determine what proposition is expressed by *If A, then B*, we diagonalize. We consider the set of context-world pairs in which the writer of the note in that context wrote down a true proposition in the world. It is easy to see that *If Pete called, he won* is true with respect to the following context-index pairs: $\langle f, w_1\rangle, \langle f, w_2\rangle$ and $\langle g, w_2\rangle$, while the assertion *If Pete called, he lost* is true with respect to the following context-index pairs: $\langle g, w_1\rangle, \langle f, w_3\rangle$ and $\langle g, w_3\rangle$. What I learn when I accept both sentences is not who made what statement, nor whether the one or the other is true, I only learn that we have to be in w_1; only in w_1 both sentences can be true. I conclude that Pete folded.[44]

There is a different, but related, way to account for Gibbard's problem without giving up the assumption that all conditionals state propositions. For all sentences whose interpretation depends on context, there are two ways in which two agents can disagree about its truth value. First, they can agree about what is said, but disagree about whether what was said is true, and second, they can have identical beliefs about the world in all relevant ways, but disagree about the truth of the sentence because they disagree about what is said. In case of conditionals the latter kind of disagreement can be accounted for by saying that the way to select nearest worlds differs. Even if I have all the relevant information of both Jack and Zack about the cards and the dispositions of Pete, there is both a way to think of *If Pete called, he lost* as being true and as being false. The conditional is true, if the similarity relation depends primarily on the cards that Pete and Mr. Stone have, because Mr. Stone has better cards. If not the cards, but the dispositions of Pete determines similarity, the conditional is false: Pete calls only if he wins. Let us say that selection function f goes with similarity by cards, and selection function g with similarity by Pete's disposition. Then the two propositions asserted by Zack and Jack are respectively $\{w \in W|\ f_w(A) \subseteq \neg B\}$ and $\{w \in W|\ g_w(A) \subseteq B\}$. Gibbard's problem is no threat to the principle of conditional non-contradiction as long as the latter is restricted to the proposition expressed in a fixed but arbitrary context, because Zack's and Jack's use of respectively the sentences *If Pete called, he won* and *If Pete called, he lost* simply do not express contradictory propositions. We saw already how to account for the fact that from their respective claims I can conclude that Pete folded.

The two ways to account for Gibbard's problem correspond with the two ways in which the meaning of '$>$' depends on context. According to the diagonalisation solution the two statements are not contradictory because the belief states of the two agents are different. According to the second solution the reason is that the propositions that potentially determine similarity are different from each other. But both proposals have the following in common: What is expressed by a conditional sentence is functionally dependent on the

[44]For the analysis I have made use of a *type*-analysis, but that is not essential. According to a token analysis we should distinguish six possible worlds. One of those worlds, in fact the only world that makes both assertions true, is the world where the belief state of the utterer of the conditional *If Pete called, he won* can be represented by $\langle\{w_1, w_2\}, f\rangle$, and the belief state of the utterer of *If Pete called, he lost* by $\langle\{w_1, w_3\}, g\rangle$.

intention of the speaker; the criteria for selecting nearest possible worlds. If we say that the intention of the speaker is the relevant contextual factor, we can say that the character expressed by *if A, then B* is $\lambda f.\{w \in W|\ f_w(A) \subseteq B\}$.

5.9 Subjunctive conditionals again

According to the Ramsey test analysis, $A > B$ is accepted in K iff B is accepted in K revised by A. The triviality results showed why this analysis is not as obviously true as it was hoped for at one time. However, the problem posed by the triviality results can, at least formally, be accounted for by making the conditional context dependent. We have seen that this was proposed by Harper. But Harper (1976b) did not claim that his analysis of conditionals should be used for all kinds of conditionals, in particular, that it should be used for the analysis of counterfactuals. He argued that the analysis should only be used for those conditionals that more or less reflect the conditional beliefs of the agents who utter them. There are various reasons to think why subjunctive conditionals should not be handled in this way. First, as observed by Adams (1970), an analysis of counterfactuals in terms of the actual conditional beliefs of the agent cannot account for certain *explanatory uses* of counterfactuals. As noted by Stalnaker (1975), there are subjunctive conditionals whose antecedents are consistent with what is presupposed, but for whose interpretation we necessarily should look outside the context that represents this common background knowledge. He suggests that this is exactly the reason why we use the subjunctive mood. In a sentence like *If Mary were allergic to penicillin, she would have exactly the symptoms she is showing* the conditional is presented as evidence for the truth of its antecedent. If subjunctive conditionals are handled via the epistemic Ramsey test analysis, and if the relevant context is that what is currently presupposed, the sentence would be trivially true and so could be no evidence for the truth of the antecedent.

The most convincing reason why subjunctive conditionals should not be analyzed via the most straightforward reading of the Ramsey test analysis is of course that it becomes unclear how we could account for the difference between Adams' Oswald-Kennedy examples when they are stated in indicative and subjunctive mood. In a similar way it becomes impossible to account for the following:

> Suppose I accept that if Hitler had decided to invade England in 1940, Germany would have won the war. Then suppose I discover, to my surprise, that Hitler did in fact decide to invade England in 1940 (although he never carried out his plan). Am I now disposed to accept that Germany won the war? No, instead I will give up my belief in the conditional. In this case, my rejection of the antecedent was an essential presupposition of my acceptance of the counterfactual, and so gives me reason to give up the counterfactual rather than to accept its consequent, when I learn that the antecedent is true. (Stalnaker, 1984, p. 105).

Let A and B be *Hitler decided to invade England in 1940* and *Germany would have won the war*, respectively. As Gärdenfors (1988) noted, to account for this example in the global epistemic approach towards revision, that is, giving up $A > B$ rather than accept B as a response of learning A, it is needed that $\neg B$ is more strongly entrenched than $A > B$. The problem is that this cannot be the case. According to the epistemic account, conditionals do not really express propositions. They are only accepted or not in a whole belief state represented by something like a system of spheres. My conclusion is that at least some counterfactuals must denote propositions. But that some counterfactuals must denote a proposition doesn't mean that they are thus context independent. We have seen already that we can say that even indicative conditionals express propositions, although what is expressed by such indicative conditional sentences is very context dependent. What is expressed by an indicative conditional is extremely context dependent because it not only depends on the speaker's criteria for selecting, but also on the particular belief state of the speaker. The proposition expressed by a subjunctive conditional sentence is not so extremely context dependent. But – as we have seen in section 5.8 – it is already possible that even if only the propositions that potentially determine nearness depends on context, two subjunctive conditional sentences of the form *If A were the case, B would be the case* and *If A were the case,* $\neg B$ *would be the case* can both be true at the same time.[45]

From now on we will say that counterfactuals are simply true or false in a world according to a contextually given selection function. If this is so, counterfactuals express propositions and can thus be less strongly entrenched than their consequent. In particular for the Hitler example, it becomes possible now that the counterfactual $A > B$ is given up because learning A, that Hitler decided to invade England in 1940, does not result in giving up my belief in $\neg B$, that Germany lost the war.

It seems we have come to the same conclusion as Lewis (1973, 1975) and Gibbard (1980): the Ramsey test is relevant for the analysis of indicative conditionals, but this is not the case for counterfactuals. But their position leaves an important question to be answered: if the selection function is not to be explained as the projection of a methodological policy onto the world, how then should we understand the meaning and role of counterfactuals?

We have seen convincing arguments why belief in counterfactuals should not be explained in terms of conditional beliefs in the most straightforward reading of the Ramsey test analysis. But that does not mean that the project of explaining the meaning of counterfactuals in terms of conditional beliefs is completely hopeless.

> If we could distinguish and filter out those aspects of our epistemic situation which derive more from our parochial perspective and less from the way we take the world to be, we might be able to explain the acceptance of conditional propositions in terms of the open conditional that would be acceptable in ide-

[45] See chapter 7 of Stalnaker (1984) for more discussion.

> alised contexts which abstract away from those aspects. (Stalnaker, 1984, pp. 115-116)

From those suggestions to an account of the connection between beliefs in counterfactuals and conditional beliefs is a long way, and I have not much to offer. It is clear what should be accounted for: the fact that we normally understand each other if we use counterfactuals. This doesn't mean that the counterfactual connective thus has a fixed interpretation, not even if the set of propositions that potentially determine similarity is fixed. It still depends on what is accepted. But this acceptance state need not be the actual belief state or presupposition state of the agent. That the use of counterfactuals does normally not lead to interpretation problems suggests that in most uses of counterfactuals it is relatively clear what criteria for selecting is assumed by the utterer. This, in turn, means that the pragmatics of conditionals must be quite systematic. It is possible that for some uses of subjunctive conditionals the selection function reflects the current belief state of the utterer, sometimes his prior belief state,[46] sometimes an information state that is simply consistent with natural laws, and sometimes something else. Thus, I believe that the meaning of the conditional connective, '$>$', should in the end be explained in terms of conditional beliefs, but the relevant belief state can be a prior belief state, or an information state that reflects the beliefs of a great number of agents, or maybe a combination of both. Of course, once it is assumed that the meaning of the connective '$>$' should be explained in terms of conditional beliefs, and if conditional beliefs are interpreted by conditionalisation or qualitative variants thereof, the results of Stalnaker (1970a) and van Fraassen (1976) might be relevant again.

The pragmatics of conditionals starts with the assumption that the selection function is context dependent. In general it is difficult for counterfactuals to say more about in what way the selection function is determined by context. One pragmatic aspect about conditionals, however, is pretty clear. This is the way the selection function changes during an argument. To that we will turn now.

5.10 Invalidity as illegitimate change of context

The meaning of a conditional depends on the way similarity is measured. If a speaker asserts a (subjunctive) conditional, he has a certain way of selecting similarity in mind. If

[46] That we should look at a prior state in one way or another (i.e. a prior belief state, or a prior state of the world) for the analysis of counterfactuals has been proposed by a number of people. Adams (1976) was probably the first, followed by for instance Thomason & Gupta (1980), Lewis (1979c) and Skyrms (1980a/b). Thomason & Gupta suggest that looking at current versus a prior state is all there is to the distinction between Adams' Oswald-Kennedy examples in respectively indicative and subjective form. Lewis (1979c) argued that the notion of similarity is not as vague as has been suggested by Fine (1975) (and later by Kratzer, 1989), if it is recognised that for determining similarity, prior states of the world are crucial. I don't believe though that looking at a prior state can be everything there is to the subjunctive mood.

the selection function used for the interpretation of counterfactuals can be dependent on the speaker's intention, a hearer can disagree in two ways with the speaker with respect to the truth value of a counterfactual. It can be the case that the hearer understood the speaker correctly and that they disagree about the facts. But, as in other cases of context dependence, it is also possible that the hearer has misunderstood the speaker's intention. He disagreed with the speaker because he assumed a different way of selecting nearest possible worlds, he picked out the wrong selection function. He misunderstood the speaker because he assumed that a different proposition was expressed to what the speaker actually wanted to express.

What is problematic about the analysis of conditionals is not only that it is difficult to determine what the relevant set of propositions is that determines similarity, but also that this set should stay stable during an argument involving more conditionals. In inferences where in the middle of the argument the set of propositions that determine similarity is changed, a fallacy will arise. Let S and S' be two sets of propositions that potentially determine similarity. Let us say that if first S and then S' measures similarity, a *context change* has occurred. In principle the set S' can stand in four kinds of relations to S; S' can be independent of S, S' can be a subset or a superset of S, and finally S and S' can be disjoint.

We have already seen one case, Gibbard's example, where S and S' do not stand in a sub- or superset relation to each other. In such a case, different things might be expressed with the same conditional sentence. In other interesting cases, S and S' *do* stand in an inclusion relation in one way or another, and the selection function that corresponds to one set is thus more fine grained than the selection function that corresponds with the other.

Let us first look at a case where the set of propositions that determine similarity decreases during the argument. In these cases a context change occurs, but we typically find it difficult to detect this change of context. Some famous fallacies typically arise in these kind of circumstances. Consider the following argument for fatalism of Dummett:

> Either I will be killed in this raid or I will not be killed. Suppose that I will. Then even if I take precautions I will be killed, so any precautions I take will be ineffective. But suppose I am not going to be killed. Then I won't be killed even if I neglect all precautions; so, on this assumption, no precautions I take will be either ineffective or unnecessary, and so pointless. (from Stalnaker, 1975)

The argument is of the following form: $K \vee \neg K$, (if K, then (if P then K), thus Q), (if $\neg K$, then (if $\neg P$, then $\neg K$), thus R), thus Q or R. The argument is invalid, because the statements *If P, then K* and *If $\neg P$, then $\neg K$* are not valid. But, as Stalnaker points out, in the contexts in which these conditionals are used (respectively K and $\neg K$), they give rise to reasonable inferences (for the notion of reasonable inference, see section 5.6.6). The problem with the argument, according to Stalnaker, is that it assumes that the conclusion is a reasonable inference given that the sub-arguments are reasonable inferences. But that

is not the case. This is only the case if all the sub arguments are reasonable inferences with respect to the *same* context, which was not the case in the fatalism argument. The conditionals used in the sub arguments are true in the contexts where respectively K and $\neg K$ are accepted. But the conditionals can no longer be accepted in the main context, a context where neither K nor $\neg K$ is accepted.

To illustrate, consider the following belief state: $\langle W, f\rangle$, where $W = \{w_1, w_2, w_3, w_4\}$, $K = \{w_1, w_2\}, P = \{w_2, w_3\}, Q = P > K$, $R = \neg P > \neg K, f_{w_1}(P) = f_{w_2}(P) = f_{w_3}(P) = f_{w_4}(P) = P, f_{w_1}(\neg P) = f_{w_2}(\neg P) = f_{w_3}(\neg P) = f_{w_4}(\neg P) = \neg P$.

It is clear that in this belief state, $Q \vee R$ is not true, so the inference is not valid. But because in $\langle W, f\rangle$ the preservativity condition is satisfied, the inference is not even reasonable. Still, the sub arguments are reasonable because if you assume K and preservativity, you end up in belief state $\langle K, g\rangle$, where $g_{w_1}(P) = g_{w_2}(P) = \{w_2\}$, and if you assume $\neg K$ and preservativity, you end up in belief state $\langle \neg K, h\rangle$, where $h_{w_3}(\neg P) = h_{w_4}(\neg P) = \{w_4\}$. Note that g and h are simply the selection functions $f^{W\cap K}$, and $f^{W\cap \neg K}$, respectively.

We have seen that in analyzing conditionals in discourse or argument, we easily go from a more to a less determined selection function. The less determined the selection function is, the more worlds in which the antecedent is true we have to check as to whether the consequent is true. Thus, the more difficult it will be for a conditional to be true.

5.11 The systematicity of context change

Lewis and Stalnaker recognised the context dependence of the selection function for the analysis of counterfactuals. What they did not so clearly see, I think, is that this context dependence is in some cases very systematic. An important argument for both was that counterfactuals of the form $A > C$ and $(A \wedge B) > \neg C$ can be true simultaneously. However, as noted by Frank (1997), only discourses of the form "$A > C$, *and* $(A \wedge B) > \neg C$" are acceptable, the same discourse in the converse order is out. But if only truth mattered, and the two counterfactuals would be interpreted via the same selection function, Lewis and Stalnaker would not predict a difference. It seems that the interpretation of a counterfactual changes the context in such a way that other kinds of counterfactuals can no longer be appropriately uttered in the new context. Frank argues that in fact the inappropriateness of a conjunction of the form "$(A \wedge B) > \neg C$ and $A > C$" calls for a variable strict conditional account. Let me first give an intuitive motivation for this account.

It seems reasonable that any adequate theory of counterfactuals must account for the fact that at least most of the time instantiations of the following formula (Simplification of Disjunctive Antecedents, SDA) are true:

$$(SDA) \quad [(A \vee B) > C] \rightarrow [(A > C) \wedge (B > C)]$$

The problem is that if we make this principle valid, by saying that $f_w(A \vee B) = f_w(A) \cup f_w(B)$, the theory looses one of its most central features, its non-monotonicity.

The principle of monotonicity,

$$(MON) \quad [A > C] \rightarrow [((A \wedge B) > C)],$$

becomes valid. That is, by accepting SDA, we can derive MON on the assumption that the connectives are interpreted in a Boolean way.[47] The Lewis/Stalnaker account does not validate MON because SDA is not a theorem of their logic. The same is true for Adams' probabilistic account. However, are those who claim that SDA should be a theorem not right? It certainly is the case that from (113a) we infer (113b) and (113c):

(113) a. If Spain had fought on either the Allied side or the Nazi side, it would have made Spain bankrupt.

b. If Spain had fought on the Allied side, it would have made Spain bankrupt.

c. If Spain had fought on the Nazi side, it would have made Spain bankrupt.

Contrary to Lewis and Stalnaker, the inferences are predicted to be valid by a strict conditional account.[48]

5.12 A variable strict conditional account

The oldest way to account for the peculiarities of counterfactual conditional was to interpret them as modalised material conditionals. Thus, if '$\rightarrow$' denotes the material conditional, $A > B$ is true in w iff $A \rightarrow B$ is true all worlds w' that are accessible from w. The strict conditional account predicts that transitivity, strengthening of antecedent, and contraposition are all valid.[49] Stalnaker (1968) and especially Lewis (1973) argued that counterfactuals cannot be analyzed as strict conditionals, because in that way we cannot account for certain fallacies. In particular, counterfactuals do not behave in a monotone way and don't obey transitivity and contraposition. And they are right: if the accessibility relation stays constant, a strict conditional account will not do. According to Lewis and Stalnaker (and Adams), we should give a *semantic* account of the fallacies associated with counterfactuals.

[47] From $A > C$ and the assumption that connectives are interpreted in a Boolean way, we can derive $((A \wedge B) \vee (A \wedge \neg B)) > C$. By SDA we can then derive $(A \wedge B) > C$.

[48] With Fine (1975) I don't think that this means that counterfactuals with disjunctive antecedents falsify the Lewis/Stalnaker account. The reason is that we cannot conclude *If A, then C* from all instantiations of conditionals of the form *If A or B, then C*: "If Spain had fought on either the Allied side or the Nazi side, it would have fought on the Nazi side. Thus, if Spain had fought on the Allied side, it would have fought on the Nazi side." (McKay & van Ingwagen, 1977).

[49] The three principles are closely related to each other (See Stalnaker, 1984): From *transitivity* to *strengthening of antecedent*: Immediate, if $(A \wedge B) > A$ is assumed to be valid. From *contraposition* to *strengthening of antecedent*: Assume weakening the consequent (if C is entailed by B, then $A > C$ is entailed by $A > B$). Suppose $A > C$, by contraposition $\neg C > \neg A$, by weakening the consequent $\neg C > \neg(A \wedge B)$, by contraposition $(A \wedge B) > C$.

However, we have seen that to account for other fallacies, both take the relevant selection function to be context dependent, and that this context dependence seems to change systematically in a discourse. But if the relevant selection function sometimes systematically has to change during an argument, does the Lewis/Stalnaker account still have an advantage over a strict conditional account if we allow the accessibility relation to change during an argument? That all depends on how the accessibility relation is defined, how it can change during an argument, and how straightforward a strict conditional analysis can account for the fallacies associated with counterfactuals.

In a very interesting article, Warmbrod (1981) argues for something like a strict implication analysis, and accounts for all kinds of fallacies related to counterfactuals by the principle that the relevant accessibility relation is not allowed to change during the argument. Remember that according to the strict implication account a counterfactual *if A then B* denotes the following proposition: $\{w \in W : \forall w' \in W[wRw' \Rightarrow w' \in (A \rightarrow B)]\}$, where R is an accessibility relation and $\rightarrow$ material implication. Equivalently, this is just $\{w \in W : R(w) \subseteq (A \rightarrow B)\}$, the selection function is replaced by an accessibility relation.

What makes Warmbrod's analysis interesting is his claim that this accessibility condition should satisfy two constraints. First, he argues that if we interpret several conditionals that intuitively 'belong together', we should analyze all those conditionals with the help of the *same* accessibility relation. If we analyze an argument to which a set of conditionals belong, we have to analyze those conditionals with respect to the same context, i.e., the accessibility relation is not allowed to change during the argument. The second requirement is that all the *antecedents* of the conditionals should be *consistent* with the set of accessible worlds. So, if we analyze a set of conditionals in a possible world w, we assume a contextually determined set of worlds, $R(w)$, with which all the antecedents of the conditionals are consistent. All cases that have been assumed to be counterexamples to transitivity, contraposition, monotonicity, and SDA are special in that there is no single accessibility relation such that all the corresponding material implications of the premises are true in all the accessible worlds to w in which the extra constraints are satisfied. It follows that a strict conditional account is not so bad as is suggested by Lewis (1973). That the (apparent) counterexamples to monotonicity can be explained away as suggested above is clear,[50] but it is also true in the case of the other three principles. Consider an (apparent) counterexample to transitivity:

(114) a. If Bush (senior) had not lost the election in 1992, Clinton would not have been President in 1994.

b. If Bush (senior) had died during the Gulf war in 1990, he would not have lost the election in 1992.

[50]When $A \rightarrow C$ is true in all accessible worlds, and intuitively also $(A \wedge B) > \neg C$ is true, there will be no accessible world in which $A \wedge B$ is true.

c. If Bush (senior) had died during the Gulf war in 1990, Clinton would not have been President in 1994.

According to Warmbrod, this would not be a counterexample to transitivity, because there is no (reasonable) set of worlds $R(w)$, such that (i) $R(w)$ is consistent with the antecedents of (114a), (114b) and (114c), and (ii) the material conditionals corresponding with (114a) and (114b) are also true in all worlds in R(w). The three sentences can only be true together, according to Warmbrod, if we change the context during the argument. But that is not allowed. In the first premise we assume that in all the worlds of the sphere suggested by the antecedent it is true that Bush was running for President in 1992, while this cannot be true in the worlds verifying the antecedent of (114b). The (apparent) counterexamples to contraposition and SDA can be explained away in similar ways.

Okay, the strict conditional analysis can account for the fallacies associated with counterfactuals. Still, two questions remain: First, where does the relevant accessibility relation come from? and second, can it give a natural explanation why discourses of the from *If A and B, then not C, but if A, then C* are so much worse then their converse *If A, then C, but if A and B, then not C*?

To the first question Warmbrod gives a rather surprising answer: When counterfactuals are interpreted 'out of the blue', the accessibility relation should be based on the Lewis/Stalnaker notion of similarity. Let $\leq_w$ be the order relation that underlies the Lewis/Stalnaker analysis of counterfactuals. He then assumes that to interpret the conditional $A > C$ out of context in world w, $R(w)$ is defined as $\{v \in W |\ \forall u \in A :\ v \leq_w u\}$. Let us call this set $R_A(w)$. Of course the Lewis/Stalnaker selection function can be easily defined in terms of $R_A(w)$ as $f_w(A) = R_A(w) \cap A$. It follows that Warmbrod's analysis only differs in an interesting way from our familiar account when a counterfactual $A > C$ is interpreted in a context in which a number of other counterfactuals have been uttered. We have already seen that the (apparent) counterexamples for the strict conditional analysis can be explained away by context change. But how does this context change work, and how systematically can we account for it?

First, note that in general $R_A(w) \subseteq R_{A \wedge B}(w)$, the requirement to be consistent with both A and B is bigger than the requirement of just being consistent with A. If we first state $(A \wedge B) > \neg C$ out of context, the relevant set of accessible worlds will be $R_{A \wedge B}(w)$. Because A is consistent with this set, for interpreting $A > C$ we don't have to change the accessibility relation. Instead, the conditional $A > C$ will simply be false according to the strict conditional account. If we state the conditionals in the inverse order, however, the first accessibility relation will be $R_A(w)$. In all apparent counterexamples to strengthening of the antecedent, $A \wedge B$ will not be consistent with $R_A(w)$. We have to consider more possible cases and so make the conditional more specific. With respect to this changed context it might very well be that the strict conditional $\Box((A \wedge B) \rightarrow \neg C)$ will be true. But the interesting question is by means of what procedure the relevant accessibility relation changes in the latter example from $R_A(w)$ to $R_{A \wedge B}(w)$?

The answer that suggests itself immediately is *accommodation*: A conditional can only be appropriately uttered with respect to a context if the antecedent of the conditional is consistent with this context. For indicative conditionals this context is what is believed or presupposed by the agent in w, while for subjunctive conditionals it is this distinguished set $R(w)$. What if this appropriateness condition is not fulfilled in w? In that case we accommodate the relevant context such that it satisfies the requirements.

I do believe that this can be worked out, but at a certain cost. Normally it is assumed that accommodation is a repair strategy that has to be used only in certain special situations. This assumption has to be given up, however, if we want to analyze conditionals as strict conditionals. The reason is that in a lot of cases we have to change (accommodate) the relevant accessibility relation before we can analyze the counterfactual as a strict conditional.[51] Fortunately, we don't have to make such heavy use of accommodation if (i) we assume that conditionals are analyzed in terms of selection functions, and (ii) we allow the relevant selection functions to change their denotations through conversational means.[52] We have already seen how this can be worked out for the analysis of *indicative* conditionals. But we have argued that this is not enough, because the relevant information state with respect to which *subjunctive* conditionals are to be interpreted need not be the actual belief or presupposition state of the agent. But how then should we account for the change of selection functions for the analysis of counterfactuals?

5.13 Change of selection function

Although in this section I am going to argue that the selection function should be able to change during the conversation, at first sight it seems as if this is not really needed at all. Once we make use of *modal subordination*, as in chapter 4, we can say that after the assertion of a subjunctive conditional of the form $(A \wedge B) > \neg C$, we introduce a modally subordinated context to the discourse, and interpret the antecedents of a later (subjunctive) conditional with respect to this subordinated context. If this later conditional has the form $A > C$, it really means $(A \wedge B) > C$ and thus states exactly the opposite of the first conditional, which is incoherent.

I believe that this modal subordination account indeed goes a long way, but unfortunately, it cannot account for the informativity of a sequence of conditionals like "*If Bush*

[51] See also Stalnaker (1984, p. 126).

[52] Nute (1984) has noted another problem for Warmbrod's strict conditional account. If counterexamples to valid inferences according to the strict conditional account are to be explained away by illegitimate context change, it seems natural that we can always make up such counterexamples if we change the context in the middle of the argument. The following principle is valid according to the strict conditional account: $[(A > C) \wedge (B > C)] \Rightarrow [(A \vee B) > C]$. Suppose now that $B \cap R_A(w) = \emptyset$ and $R_A(w) \subseteq R_B(w)$. In that case we would change the context in the middle of the argument. But in this changed context it is not at all necessary that $R_B(w) \cap A \subseteq C$, and thus that $R_B(w) \cap (A \vee B) \subseteq C$. However, it seems hard to find counterexamples to the above principle.

had not lost the election in 1992, Clinton would not have been President in 1994. And if Bush had died during the Gulf war in 1990, he would not have lost the election in 1992." We can conclude that our account has to be a bit more complicated.[53]

Given the way Harper's analysis of iterated revision works, it is quite easy to account for the systematic change of selection functions. Let us fix a set S of propositions that, by Harper's principle, potentially determine similarity. The actual similarity relation then not depends only on set S, but also on which propositions of S are decided by a particular context, K. We have seen earlier that by Harper's principle, $C''_K(A)$ will be a subset of K, if A is consistent with K, and picks out a set of A-worlds that are very close to K-worlds otherwise. Note now that if f^K is a selection function, $f^{C''_K(A)}$ is a selection function too. Let us now assume that any selection function f^K is interpreted like f, but is preservative with respect to K. Thus, for any w, K, and A, $f^K_w(A)$ is defined as follows:

$$\begin{aligned} f^K_w(A) &= f_w(A \cap K), \text{ if } A \cap K \neq \emptyset \\ &= f_w(A), \text{ otherwise.} \end{aligned}$$

If $g = f^K$, let us say that $g^A = f^{C''_K(A)}$. Let us now assume that at the beginning of a conversation, any selection function of a possibility with world w is of the form $f^{\{w\}}$. Note that even if $A \cap B \neq \emptyset$, then $C''_{\{w\}}(A)$ need not be consistent with B, which is enough to get rid of the problem that plagued the modal subordination account suggested above.[54] Let us now assume that after the interpretation of the counterfactual $A > C$ in possibility $\langle w, f\rangle$, the selection function relevant for the analysis of conditionals changes from f to f^A. If we then would follow this conditional with the (subjunctive) conditional $B > D$ with respect to this changed possibility, $\langle w, f^A\rangle$, the latter conditional would then have the same truth value as $(A \wedge B) > D$ would have in possibility $\langle w, f\rangle$, if B is consistent with the relevant set of selected A-worlds.

But this is exactly what we need to explain the difference between the acceptable discourse "$A > C$, *but* $(A \wedge B) > \neg C$" on the one hand, and the unacceptable "$(A \wedge B) > \neg C$, *but* $A > C$" on the other; the second conditional of the second discourse is interpreted with respect to selection function $f^{A \wedge B}$, if the first conditional was interpreted with respect to f, and thus states exactly the opposite of what is asserted by the first conditional, which is incoherent.

Note that because we allow our selection function to change during the discourse, we can account for the coherence, or even validity, of discourses like *If A were the case, then C were the case. If A and B were the case, then C would not be the case..... If A were the case, then C were not the case.*

To sketch how our proposal can be implemented, I will give a simultaneously recursive definition of *truth* and *context change* for a simple propositional language as before, where

[53] Although you might argue that these counterexamples are examples that should not be accounted for by means of modal subordination.

[54] Consider B to be the proposition that Bush has lost the election in 1992.

I make the simplifying assumption that all conditionals are subjunctively used, that the antecedent of a conditional is always atomic, and that the selection function is the only relevant part of the context. The truth definition is simple, as always. The only interesting clauses are the ones for conjunction, and the subjunctive conditional:

- $[[A > B]]^{w,f} = 1$ iff $f_w(\{v \mid [[A]]^{v,f} = 1\}) \subseteq \{v \mid [[B]]^{v,Upd(A,\langle w,f\rangle)} = 1\}$
- $[[A \wedge B]]^{w,f} = 1$ iff $[[A]]^{w,f} = 1$ and $[[B]]^{w,Upd(A,\langle w,f\rangle)} = 1$

Note that I assume that the selection function with respect to which the consequent and second conjunct are interpreted need not be the same as the selection function w.r.t which the whole assertion is interpreted. If this selection function is g^K, the selection function relative for the consequent will be different if the antecedent is not entailed by K, and the selection function relative for the second conjunct will be different if the first conjunct contains a subjunctive conditional that satisfies the above requirement.[55] How the selection function changes is defined as follows:

- $Upd(A, \langle w, f\rangle) = f$, if A is atomic
- $Upd(\neg A, \langle w, f\rangle) = Upd(A, \langle w, f\rangle)$
- $Upd(A \wedge B, \langle w, f\rangle) = Upd(B, \langle w, Upd(A, \langle w, f\rangle)\rangle)$
- $Upd(A > B, \langle w, f\rangle) = Upd(B, \langle w, f^A\rangle)$[56]

5.14 Conclusion

I have shown how we can analyze conditional sentences as statements that express propositions, once we take context dependence seriously. Which proposition a conditional sentence expresses might depend on a relevant information state and how it behaves under revision. In this chapter I have discussed the widely held assumption that conditional sentences should be analyzed in terms of belief revision. In the next chapter I will argue that also for a large number of attitude attributions the analysis of belief revision is important.

[55] The same will be true if we define inferences between sequences of sentences in the following way: C can be inferred from $A_1, ..., A_2$ iff for all $\langle w, f\rangle \in W \times F$: if $[[A_1 \wedge ... \wedge A_n]]^{w,f} = 1$, then $[[A_1 \wedge ... \wedge A_n \wedge C]]^{w,f} = 1$.

[56] I am not sure whether I should change this update rule in case the antecedent itself is, or can contain, a subjunctive conditional. Therefore I stick to the simplifying assumption that this won't happen.

Chapter 6

Some other attitudes

6.1 Introduction

In the previous chapter I discussed conditionals and belief revision. One of the results of that discussion was that belief states should not simply be represented by a set of possible worlds, but rather by an ordering relation, or, equivalently, by a set of worlds plus a change function. I argued that the more common kind of change function, or revision policy, depends to a large degree on what the agent accepts. Still, a state that is represented by a set of worlds plus a change function contains more information than the state just represented by the set of worlds alone.

In this chapter I argue that the extra information that is represented by such a belief, or information state, is useful to analyze a number of other attitudes than *belief*. The reason is that in this way we can make a difference between different propositions believed: proposition A is more strongly believed than proposition B, because it is more strongly connected with other beliefs than proposition B. A number of *evidential verbs*, for instance, will be analyzed in terms of *robustness* under belief revision. I argue that this richer representation of belief states will also be useful for the analysis of attitudes of desire. In particular, that *intention* is a strong kind of desire: you intend something if you not only desire it in your present belief state, but the desire is also *robust* under belief revision.

We have also seen in the previous chapter that it is good to distinguish two kinds of belief change: (i) the change of belief due to receiving more *information*, and (ii) the change of belief due to the fact that a certain *act* has taken place. In this chapter I argue that the latter kind of belief change might also be important for the analysis of certain attitudes of desire; for those attitudes where the agent has to consider the consequences of his own actions.

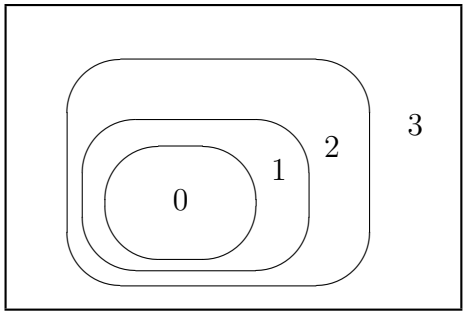
Not only belief revision per se, but also the structure used for the analysis of belief revision seems important for the analysis of desire attributions. For instance, we might want to analyze desires in terms of a *preference* order, and this preference order has a lot in common with the ordering relations underlying belief revision. Also probability and

decision theoretic frameworks, partly discussed in the foregoing chapter, will be relevant for the analysis of verbs of desire. In fact, I will argue that for a desire attribution to be true, both the preference order *and* the strength of belief are relevant.

6.2 Evidential attitudes and plausibility

6.2.1 Plausibility

The entrenchment relation used for the analysis of belief revision gives rise to a notion of *plausibility*. First, it gives rise to a plausibility grading of the possible worlds. On the basis of an information state K and a set of propositions S that potentially determines similarity (cf. section 5.7), we determined in the previous chapter a *qualitative* ordering relation on possible worlds $\preceq_K$. But given that we *count* the propositions in S, we can even define a more informative *quantitative* function k, $k(w/A) \stackrel{def}{=} |S^w_u C''_K(A)|$, for any u in $C''_K(A)$. The measure $k(w/A)$ represents the plausibility of w after revising the information state K with A. The idea is that $k(w/A)$ is the number of propositions decided by $C''_K(A)$ that potentially determine similarity on which w and any arbitrary element of $C''_K(A)$ differ in truth value. The higher $k(w/A)$ is, the less plausible the agent in belief state k would find world w after he would revise his belief state by A. The function k represents an *extended* belief state; it represents not only what the agent believes, $K \stackrel{def}{=} \{w \in W : k(w/\top) = 0\}$,[1] but also how plausible he considers worlds outside of K. We can illustrate this by the following picture, where the numbers indicate the k-value of the worlds in the ovals:



In terms of the (conditional) plausibility of worlds, we can determine the (conditional) plausibility of propositions, i.e in terms of $k(w/A)$ we can define $k(B/A) \stackrel{def}{=} min\{k(w/A) : w \in B\}$. The measure $k(B/A)$ represents the degree of disbelief in B given that A is true. If $k(B/A) = 0$, this means that B is consistent with the belief state resulting after revision of our current belief state with A. For those who have seen Spohn's (1987), it is obvious that the above plausibility functions are simplified versions of his ordinal conditional functions. Let us abbreviate $k(A/\top)$ by $k(A)$. Then we can follow Spohn in saying that A is accepted

[1] Above we defined k in terms of K (plus a set of propositions S), but we can also go the other way round, take k to be primitive, and derive K from it.

in k, (or in K) iff $k(\neg A) > 0$, i.e., iff $\neg A$ is inconsistent with what the agent believes.[2] What k measures is potential surprise. In general, we can say that A is believed to be more plausible than $B, A > B$, iff $k(\neg A) > k(\neg B)$ or $k(A) < k(B)$. Given the close relation between our entrenchment relation and Spohn's ordinal conditional functions, it should not come as a surprise that our entrenchment relation satisfies the five Gärdenfors postulates (1988, pp. 88-91) for entrenchment: For all $A, B, C, K \subseteq W$:

$(EE1)$ if $A \leq_K B$ and $B \leq_K C$, then $A \leq_K C$,

$(EE2)$ if $A \subseteq B$, then $A \leq_K B$,

$(EE3)$ for all $A, B \supseteq K, A \leq_K A \cap B$ or $B \leq_K A \cap B$,

$(EE4)$ if $K \neq \emptyset$, for all $B \supseteq K, K \not\supseteq A$ iff $A \leq_K B$, and

$(EE5)$ if $B \leq_K A$ for all B, then $A = W$.

The reason that the entrenchment relation satisfies these postulates is that the following equation holds: $A \leq_K B$ iff $k(\neg A) \leq k(\neg B)$ and that it is well known (see Gärdenfors, 1988, section 4.6) that Spohn's ordinal conditional functions generate an entrenchment relation that satisfies (EE1)-(EE4) and that (EE5) follows from possible world semantics.

Let us now assume that a belief state should not be represented by a set of possible worlds, but rather by a plausibility function. It seems reasonable to assume that once we have such a richer representation of a belief state, we can account for more attitudes in terms of belief states than would be possible without such a representation. Indeed, this is what I will assume.

6.2.2 Evidential verbs

It seems reasonable that verbs like *be certain*, *be sure*, *be convinced* and the future looking *expect*, and *predict* should be analyzed as *believe* + some extra condition. The reason is that from *a is sure that A* we can conclude that *a* believes that *A*. What should this extra condition be? Following Asher (1987), it should at least guarantee the principles of *belief inference*, (B); *simplification*, (S); *conjunction introduction*, (I $\wedge$); and *upward entailment*, (UE), for all these kind of verbs (where α is the attitude verb):

B:	a αs that A	$\Rightarrow$	a believes that A
S:	a αs that $(A \wedge B)$	$\Rightarrow$	a αs that A
I$\wedge$:	a αs that A & a α that B	$\Rightarrow$	a αs that $(A \wedge B)$
UE:	a αs that A & $A \subseteq B$	$\Rightarrow$	a αs that B

[2] An alternative way to define K in terms of k is thus as follows: $K \stackrel{def}{=} \bigcap\{A \subseteq W |\ k(\neg A) > 0\}$.

The extra condition for these verbs is *evidential* in nature and should be some kind of *justification condition*. The simplest way to go about it is to assume a new accessibility function. In this way principles (S), (I$\wedge$) and (UE) follow immediately. To account for (B), this new evidential accessibility function assigns to each world w a set of worlds that is a superset of $K(a, w)$, the doxastically accessible worlds.

Although this kind of rule will do to account for the above inferences, it is preferred to account for these principles by using primitives we use already, or by primitives that are also useful for the analysis of other attitudes. I propose to account for the inferences by using the extra *inductive* information represented in a belief state if we take the notion of *epistemic entrenchment* seriously. If α is an evidential verb, *a αs that A* is true only if (i) a believes that A, and (ii) A is highly entrenched in a's belief state. In other words, it should be the case that A is believed, and that $\neg A$ is very implausible, or that A is very strongly believed. We have seen above that given a set of propositions that potentially determines similarity, a belief state K gives rise to an ordinal function, k, that measures implausibility, which in turn, via the Shackle identity, $f(A) = k(\neg A)$, gives rise to a belief function, f, that measures plausibility or epistemic entrenchment.[3] Let us say that f^a_w is the belief function associated with a in w. So my proposal comes down to the following:

$$[[\text{a } \alpha \text{ that } A]]^w = 1 \quad \text{only if} \quad f^a_w(A) \text{ is high}$$

What *high* means is context dependent, but the number should be at least bigger than 0. Note that if $f^a_w(A) > 0$, then $k^a_w(\neg A) > 0$, and thus $K(a, w) \subseteq A$. The proposed account predicts at least that from the truth of *a αs that A* we can infer that A is believed by a. Let us now see whether it also can account for the other inferences. Note first that simplification is a special case of upward entailment. Thus, if the above definition accounts for (UE), it also accounts for (S). We know already that if K is the belief state corresponding with k, and if we defined the entrenchment relation $\leq_K$ between propositions by $k(\neg A) \leq k(\neg B)$ iff $A \leq_K B$, then the entrenchment relation will satisfy the Gärdenfors postulates for (EE1)-(EE5). In particular it satisfies (EE2), if $A \subseteq B$, then $A \leq_K B$, and this is enough to guarantee that our interpretation rule for evidential attitudes accounts for (UE) and thus for (S). Conjunction introduction also follows from our interpretation rule. Note that by (EE3) if both A and B are believed, then either $A \leq_K A \cap B$, or $B \leq_K A \cap B$. Thus either $f^a_w(A) \leq f^a_w(A \wedge B)$ or $f^a_w(B) \leq f^a_w(A \wedge B)$. But if both $f^a_w(A)$ and $f^a_w(B)$ are high, then also $f^a_w(A \wedge B)$ must be high, and thus conjunction introduction, (I$\wedge$), also holds for evidential attitude verbs.

By the way I interpreted evidential attitudes, these attitudes have the properties of acceptance attitudes. An acceptance attitude is an attitude that can be modelled by

[3]See Spohn (1987) who refers back to Shackle (1961). For the use of belief functions, entrenchment orderings, and belief revision in non-monotonic logic, see for instance Gärdenfors and Makinson (1994). For the relation between entrenchment orderings and non-standard probability functions, see Spohn (1987) and McGee (1994).

an *acceptance state*. An acceptance state is a consistent set of propositions closed under conjunction and implication. By modelling propositions as sets of possible worlds, the intersection of an acceptance state gives rise to a set of possible worlds. How can we arrive at this set of worlds from the above interpretation rule of evidential attitudes? Above we have assumed that *a is certain that A* iff $f^a_w(A)$ is high. Let us now say that n is the minimum of the high numbers. Remember that via the Shackle identity, $f^a_w(A) = k^a_w(\neg A)$, and that $k(A) = min\{k(w) : \ w \in A\}$, where $k(w) = k(w/\top)$ and $\top$ is a tautology. The evidential accessibility function, EVI, can now be determined in the following way:

$$EVI(a, w) \quad = \quad \{w' \in W : \ k^a_w(w') \leq n\}$$

6.2.3 Knowledge

Just like other evidential verbs, knowledge is also normally analyzed as something like *belief* plus something extra. One extra thing is obviously that what is known also has to be *true*. But true belief cannot be enough, as can be illustrated by the following examples of Russell:[4]

> It is clear that knowledge is a subclass of true beliefs. [...] There is a man who looks at a clock when it is not going, though he thinks that it is, and who happens to look at it at the moment when it is right; this man acquires a true belief as to the time of day, but cannot be said to have knowledge. There is the man who believes, truly, that the last name of the prime minister in 1906 began with a B, but who believes this because he thinks that Balfour was prime minister then, whereas in fact it was Campbell Bannerman. (Russell, 1948, pp. 170-171)

What should the extra condition be that turns a true belief into knowledge? Ramsey (1931) argued that the item of belief should be obtained by a *reliable* process, that there should be a *causal* relation between the object of knowledge and the relevant belief. Reliable processes speak about the reliability of the *channels* (instruments) by which the agents acquire their beliefs: if the channels do not function in the way it should, the normal conditions do not hold and the process is not reliable (cf. Dretske, 1981). This suggestion is obviously close to the causal information theoretic account of belief defended in chapter 1. There it was argued that John believes A iff his state of belief *indicates* that A is the case, where the notion of 'indication' was explained in terms of counterfactual dependencies and normal conditions. The suggestion now would be that the rather *general* relation between the state of the world and the internal state of the agent for the case of belief, would be replaced by a somewhat more *specific* causal relation between the fact, or information, known and the internal state. In slightly different terms, both knowledge and

[4]See also Gettier (1963) for some similar examples.

belief are indication relations, analyzed in terms of normal conditions, but for knowledge the constraints on these normal conditions are more stringent than for belief.[5]

But how does this externalistic explanation of reliability relate to the more *internal* notion of 'justified belief' used for the analysis of other evidential attitudes? It is certainly the case that we feel more justified in believing an item if we acquired this item in a reliable way. But does Russell's man not feel he has a justified belief about the time of the day when he looks at the clock? Perhaps he does, but no longer when he learns that the clock was not going; in that case he would probably give up his belief about the time of the day.

It seems that if we want to analyze knowledge as justified true belief, where justified belief is analyzed in terms of extended belief states, the 'internal' notion of justified belief and the 'external' notion of truth should somehow be related to each other. The question is *how*? I would like to propose we simply follow Hintikka's prime intuition about knowledge:

> It may be useful to remember that for us the primary sense of "I know that p" is the one in which it is roughly equivalent to "p, and no amount of further information would have made any difference to my saying so". (Hintikka, 1962, p. 52)

What this quotation suggests is that an item is known iff the item is believed, and it would not be given up by the acceptance of any new proposition that is true. Thus, an item of belief counts as knowledge, iff it is *robust* with respect to the truth (cf. Stalnaker, 1996). If we assume that P is the set of propositions that are true in the actual world, w, we can formalize this idea as follows:

$$[[know(a, A)]]^w = 1 \quad \text{iff} \quad \{v \in W : \exists B \in P : k^a_w(v/B) = 0\} \subseteq A$$

It is easy to see that this interpretation rule accounts for the *factivity* of knowledge. The reason is that one of the propositions of P will be the maximal proposition that is only true in w, the proposition $\{w\}$. It will obviously be the case that for any k and w it holds that $k^a_w(w/\{w\}) = 0$.

Let us see whether our analysis can account for Russell's problems. We have already suggested that our analysis can account for the first problem: if the agent hears that the instrument on which he based his belief was not reliable, he probably wouldn't believe anymore the item he actually believed. The second problem is also straightforwardly accounted for: if the man is informed that the late prime minister is not Mr. Balfour, he probably wouldn't believe anymore that the name of the late prime minister began with a 'B'. Notice, furthermore, that in distinction with purely causal accounts, our analysis of knowledge can also account for the fact that you might have knowledge about the future,

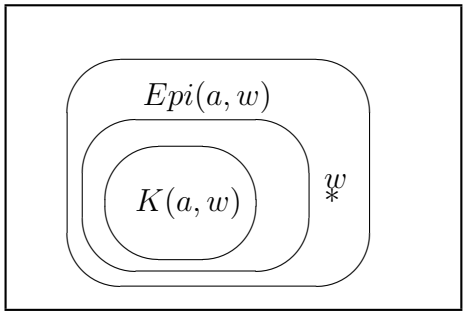
[5]Thus, the notion of 'belief' is a more *stable* attitude than knowledge, in at least some sense of the word, because defined in terms of more general relations between world and internal state than the notion of 'knowledge'.

and knowledge about so-called non-accidental disjunctive propositions like *John knows that someone was born yesterday*, where the indefinite was used non-specifically.

Another question is how the analysis of knowledge is related to our analysis of the evidential verbs discussed above. Is there any number n such that we can define an epistemic accessibility function $EPI(a, w)$ as $\{w' \in W : \ k^a_w(w') \leq n\}$? It turns out that for each belief state k^a_w there is exactly one such an n, namely the number $k^a_w(w)$. In other words, the epistemic accessibility relation can be defined in the following way:

$$EPI(a, w) \quad := \quad \{v \in W : \ k^a_w(v) \leq k^a_w(w)\}^{6,7}$$

In a picture, this would look as follows:



Obviously, if we want to say that Anton knows A in w iff $EPI(a, w) \subseteq A$, it has to be the case that $EPI(a, w)$ can also be defined as $\{v \in W| \ \exists B \in P : \ k^a_w(v/B) = 0\}$. It is easy to see that this is indeed the case.[8]

It is normally assumed that belief states are *fully introspective*; it is assumed not only that if a believes something, he also believes that he believes it, but also that if he does *not* believe something, that he believes that he doesn't believe it. This is accounted for by assuming that for every $w' \in K(a, w)$ it holds that $K(a, w') = K(a, w)$. The question that raises itself is whether also knowledge is introspective, and thus whether it should follow from our definition of $EPI(a, w)$ that for every $w' \in EPI(a, w)$ it holds that $EPI(a, w') = EPI(a, w)$.

[6]It should be noted that $EPI(a, w)$ can also be defined without making use of the *quantitative* information of k, but just in terms of the ordering relation $\preceq$ that underlies (R^*1) - (R^*4) as follows: $EPI(a, w) \stackrel{def}{=} \{v \in W : \ v \preceq w\}$. See Stalnaker (1996) for more on this. I should note that my analysis of knowledge is the same as this analysis given by Stalnaker (1996), although some of the main ideas were developed independently.

[7]Notice that if we would demand that for the analysis of evidential verbs, the number that determines $EVI(a, w)$ is higher than or equal to $k^a_w(w)$, we would assume that all evidential verbs would be *factive*. Because we don't want that, we can conclude that the relevant number in these cases should be lower than $k^a_w(w)$.

[8]**Proof:** Because each true proposition, A, is a superset of $\{w\}$, it is obviously the case that $k^a_w(A) \leq k^a_w(w)$. But this means that for each v in $\{v \in W| \ \exists B \in P : \ k^a_w(v/B) = 0\}$ it holds that $k^a_w(v) \leq k^a_w(w)$, which shows the equation from right to left. For the other side, let v be a world such that $k^a_w(v) \leq k^a_w(w)$. But then it will be the case that there is a true proposition A such that $k^a_w(v/A) = 0$, namely $A = \{v, w\}$.

Philosophers have long decided that knowledge should *not* be introspective. In particular, it doesn't seem to be the case that if one doesn't know something, one also knows that one doesn't know it. The intuition behind this decision is that whether an agent knows something or not depends partially on something *external* to the agent; something that cannot be discovered by thinking about one's own thoughts. It follows that if we want to formalize knowledge in terms of the function EPI, the function should not guarantee that for each world $v \in EPI(a, w) : \ EPI(a, v) = EPI(a, w)$.

Note that if the actual world, w, is not consistent with what a believes in this world, it will hold for all worlds v consistent with what a believes in w, $\{v \in W : \ k^a_w(v) = 0\}$, that $k^a_w(v) < k^a_w(w)$. If it were the case that for each world v where $k^a_w(v) < k^a_w(w)$ it would also be the case that $k^a_v = k^a_w$, it would immediately follow by our definition of EPI that w won't be an element of $EPI(a, v)$. As a result, it would not be the case that EPI denotes an introspective function, just as we desired. But is it the case that for each belief-world v it holds that $k^a_v = k^a_w$, and if so, why?

It is very natural to answer the first of the above questions positively, and there actually exists a very natural explanation for this.[9] We have assumed all the time that k^a_w represents the *extended* belief state of a in w. It represents not only what he actually believes in w, but also his belief revision policies. It will be the case that for all worlds in $\{v \in W : \ k^a_v = k^a_w\}$ the beliefs and belief revision policies of a will be the same as in w; we might call each such v *subjectively indistinguishable* for a from w.[10] Our question was whether we should assume that for all belief-worlds v of a in w it is the case that $k^a_v = k^a_w$. If we could argue that v is one of those worlds subjectively indistinguishable from w for a, our issue would be settled. But this argument is straightforward; if v is a belief-world, the introspectiveness of K assures already that a has the same beliefs in v as he has in w. On the assumption that the belief revision policy is determined primarily by what one believes, it is only natural to assume that the belief revision policies of a in the two worlds will also be the same. And this is enough to make sure that $k^a_v = k^a_w$.[11]

Notice that if we assume that for all v such that $k^a_w(v) = 0$ it holds that $k^a_v = k^a_w$, it follows immediately that if one knows A in w, one also believes that one knows it. And indeed, this seems a reasonable assumption to make. Another consequence of the above reasoning is that also evidential attitudes are not guaranteed to behave in a fully introspective way. This would only be guaranteed to be the case if we assumed that all $v \in EVI(a, w)$ are subjectively indistinguishable from w.

---

[9] cf. Stalnaker, (1996).

[10] The accessibility function of subjective indistinguishability, $SI(a, w) = \{v \in W : \ k^a_v = k^a_w\}$, cannot only be found in Stalnaker (1996), but was discovered independently also by Zimmermann (1999), although used for a somewhat different purpose.

[11] Our analysis of *knowledge* gives rise to the logic $S4.3$, if it is assumed that $\forall v : \ k_w(v) \leq k_w(w) \rightarrow k_v = k_w$, and characterised by an accessibility relation that is reflexive, transitive and connected, thus *not* euclidean.

According to our analysis, if I know, or justifiably believe, A, and B follows from A, it should also be the case that I know/justifiably believe B, too. A straightforward *sceptical* argument suggests that this cannot be right. John knows that there is a computer in his room, and thus, the corresponding knowledge attribution would be true. But a computer is not a holographic image of a computer. Does John thus know that the thing in his room is not a holographic image of a computer? He cannot rule out the possibility that the thing is a holographic image of a computer, so – according to the argument – the knowledge attribution *John knows that the thing in his room is not an image of a computer* is not true. But this contradicts our assumption that knowledge is closed under logical consequence.

There are (at least) two possible ways to react to this puzzle: the die-hard *sceptic*, who wants knowledge to be infallible, will conclude that thus John did not even know that there was a computer in his room. A *fallibilist* like Dretske (1970), on the other hand, seeks to explain the truth of the first knowledge attribution and falsity of the second in terms of the *context dependence* of attitude attributions. Knowledge attributions, according to Dretske are essentially contrastive, where the contrast class, or set of alternative possibilities, is contextually given. For the attribution *John knows that A* to be true in a particular context, John has to eliminate all contextually given possibilities where A is not true. The relevant alternatives with respect to which knowledge attributions are evaluated are normally possibilities consistent with what *we* take for granted, i.e. don't call into question. In this way we can explain why John knows that there is a computer in his room. However, at the moment that we mention the possibility that the thing we took to be a computer is really a holographic image of a computer, the set of possibilities relative to which the knowledge attribution is evaluated *changes*, and will contain also worlds where the thing is not a computer. As a result, the second knowledge attribution will not be counted as being true. Dretske concludes that knowledge is closed under implication, but only *with respect to the relevant alternatives.*[12]

The attentive reader has noticed already that we have seen a similar argument before. In chapter 1 we explained how the belief attribution *Bert believes that there is water in the bathtub* can both be true, and *about* water, i.e. H_2O, although Bert can't make the distinction between Earth, where 'water' denotes H_2O, and Twin Earth, where 'water' denotes the stuff with the chemical structure of XYZ. The reason was that for a belief attribution like *Bert believes that there is water in the bathtub* to be true and about H_2O, it doesn't have to be the case that Bert would have eliminated all imaginable possibilities where it is not H_2O that is in the bathtub, but only all the *relevant* possibilities; where a relevant possibility is a possibility compatible with what *we* presuppose. Thus, if we presuppose, and not call into question, that 'water' denotes H_2O, we will normally only consider alternatives where 'water' indeed denotes H_2O.

[12]See also Stine (1976) and Lewis (1996), and compare this with the evaluation of conditionals in changing contexts in the previous chapter. Lewis (1996) has stressed the difference between his position, that knowledge is closed under implication, and Dretske's, but I only see two sides of the same coin.

According to the causal, information-theoretic account of content, the content of belief and knowledge states is defined in terms of counterfactual dependencies and normal conditions. I have argued in the beginning of this section that the difference between knowledge and belief can be explained in terms of the different constraints these normal conditions have to fulfill. For knowledge attributions, the constraints will be more stringent; we will call more normal, or channel conditions into question than for the analysis of belief attributions. In the last part of this section we have argued that we should not exaggerate; if for a knowledge attribution we would always call all possible channel conditions into question, i.e. take all imaginable possibilities into account, we would never be able to escape the sceptic conclusion.

6.2.4 Be surprised

Another attitude verb that is naturally interpreted in terms of an entrenchment relation is the verb *be surprised*. We will guide our investigation again by the principles it should obey. Contrary to evidential attitude verbs, *be surprised that* is not closed under implication. If John is surprised that it snows, he need not be surprised that it rains or snows. According to Asher (1987), *be surprised* is a negative factive and the interpretation rule for those verbs should obey *factivity*, (F), *belief inference*, (B), *negation*, (N), *weakened simplification*, (WNS), and *weakened downward entailment*, (WDE):

F:	a αs that A	$\Rightarrow$	A is true
B:	a αs that A	$\Rightarrow$	a believes that A
N:	a αs that $\neg A$	$\Rightarrow$	$\neg$(a αs that A)
WNS:	a αs that $(A \vee B)$ & a $Bel(a, A \vee B)$	$\Rightarrow$	a αs that A
WDE:	a αs that A & $B \subseteq A$ & $Bel(a, B)$	$\Rightarrow$	a αs that B

That *be surprised that* should obey factivity is clear. Asher argues that *be surprised* should obey (B) because it is incoherent to say *Fred is surprised that John runs, but he doesn't believe it*. The inferences (F) and (B) should be *presuppositional* inferences, because these inferences are normally preserved under negation. In other words, we can infer from either an affirmative sentence like *Mary was surprised that John didn't get an A* or a negative sentence like *Mary was not surprised that John didn't get an A* that John didn't get an A, and that Mary believed that John didn't get an A.

I noted above that it is natural to interpret *be surprised that* in terms of epistemic entrenchment, this is suggested by the fact that in artificial intelligence research, it is common to say that if A is believed, $k(\neg A)$ is called the *surprise value* of A. The most natural interpretation of *a is surprised that A* would be that A is true, a believes A, and in the belief state before it was learned that A, it was expected that $\neg A$. Thus, in this earlier belief state $k(A)$ was high.

However, just as subjunctive conditionals are not always interpreted with respect to a *prior* belief state (or objective state of affairs), it doesn't seem to be the case that to be

surprised that A, I have had to expect $\neg A$ in a prior belief state. First, according to some schools of philosophy, the real way to be a philosopher is to be surprised by things that you have always taken for granted. Second, suppose that someone learns a mathematical theorem at a young age, and only after learning much more about mathematics sees how deep the theorem really is, i.e. how surprising the truth of the theorem is given everything else he knows about mathematics at his current state. This suggests that we should not always look at an earlier belief state, but sometimes must be able to interpret *being surprised that* in terms only of the *present* belief state. To account for these latter cases, cases of surprised$_2$, my proposal would be the following:

$$[[surprised_2(a, A)]]^w = 1 \quad \text{iff} \quad f^a_w(A) > 0 \text{ but low}$$

Note first that this interpretation rule for *being surprised that* does not predict that it is closed under implication. It is easy to imagine a situation where A is not strongly entrenched, but $A \vee B$ is, because B is. If B is strongly entrenched, $f^a_w(B)$ will be high, and thus $f^a_w(A \vee B)$ will also be high. It follows that *being surprised that* will not be closed under logical implication. As in the case of evidential predicates, it follows from being surprised that A that the agent also believes A, because $f^a_w(A) > 0$ iff $k^a_w(\neg A) > 0$. The principle of negation follows immediately from this definition. If a is surprised that $\neg A$, then it should also be the case that $\neg A$ is believed by a, in which case a cannot be surprised that A. Now we have to show that weakened negative simplification and weakened downward entailment are obeyed. Note that (WNS) is a special case of (WDE), so it is enough to show that (WDE) holds. But this follows immediately from the interpretation rule. If $f^a_w(A) > 0$, but low, and $B \subseteq A$, then via (EE2) and the Schackle identity, $f^a_w(B) \leq f^a_w(A)$. Because it is also assumed that B is believed, also $f^a_w(B) > 0$. It follows that if $f^a_w(A) > 0$ but low, it is also the case that $f^a_w(B) > 0$ but low, and thus that it is also surprising that B.[13]

6.3 Doubt

The interpretation of *doubt that* should be such that it is not closed under implication, but instead obeys *addition*, (A), *negative simplification*, (NS), and *downward entailment*, (DE):

A:	a doubts that A	$\Rightarrow$	a doubts that $(A \wedge B)$
NS:	a doubts that $(A \vee B)$	$\Rightarrow$	a doubts that A
DE:	a doubts that A & $B \subseteq A$	$\Rightarrow$	a doubts that B

[13]Ede Zimmermann (personal communication) has given the following potential counterexample to the proposed analysis: "Suppose Ede meets a friend in the street whom he has believed to be far away (or dead) and convinces himself that it is really her, then Ede would still be absolutely convinced yet at the same time surprised that she is there – at least before he learns the explanation." Is this a counterexample? Perhaps not, if being convinced and being disposed to believe it are different sets of propositions that potentially determine similarity.

Note that (A) and (NS) are both special cases of (DE); so, if we can give an interpretation rule that accounts for (DE) we seem to be ready. Ignoring anaphoric relations,[14] these data suggest that *a doubts that A* should be analyzed as *a doesn't believe that A*:

$$[[doubt(a, A)]]^w = 1 \quad \text{iff} \quad K(a, w) \not\subseteq A$$

Notice that according to this kind of interpretation rule, the following inference (Asher, 1987) is predicted to be valid:

(115) a. Fred doubts that either Mary or Alfred went to school.

b. So Fred doubts that Mary went to school and he doubts that Alfred went to school.

More generally, we predict that *downward entailment* is valid, as desired. But while this interpretation rule for *doubt that* gets the above inferences right, it is doubtful that *doubt that* actually means the same as *doesn't believe that*. Intuitively, *doubt that A* seems to mean something more like 'doesn't believe that A and his belief justifies 'not A'.[15] How should we interpret this justification condition? Given our discussion above of evidential attitudes and of *being surprised that*, it will be no surprise that I will propose the Shackle-Spohn plausibility functions again. The interpretation rule for *doubt* will then be:

$$[[doubt(a, A)]]^w = 1 \quad \text{iff} \quad f^a_w(\neg A) \text{ is high}$$

[14]With respect to anaphoric relations, Asher (1987) notes that indefinites used under belief attributions (and indefinites used in the main context) can be picked up by anaphoric expressions in the scope of *doubt that*, but indefinites in the scope of *doubt* cannot, in general, figure as the syntactic antecedents of anaphoric expressions:

(ia) John believes that *a woman* broke into his apartment.
(ib) He doubts that *she* left some fingerprints.
(iia) John doubts that *a woman* will marry him.
(iib) *He believes *she* will be unhappy.
(iiia) John doubts that a woman will marry him.
(iiib) *He doubts she will be happy.

Sometimes, however, indefinites in the scope of *doubt* can be picked up by an anaphoric expression in the scope of *believe*:

(iva) John doesn't doubt that *a woman* broke into his apartment.
(ivb) He believes that *her* perfume was unmistakably Channel No. 5. (Asher, 1987)

These anaphoric data suggest that we can refer back to indefinites in the scope of *doubt* only by making use of a *descriptive pronoun*.

[15]That is, if we ignore the intuition that doubting seems to involve active thinking – although it seems that *doubt that* can describe not just acts of doubt but also states of being doubtful, as in *Ede doubts that I will ever finish this thing.*

This interpretation rule for *doubt that* is very strong. It says that *a doubts that A* is true iff a strongly believes that $\neg A$. Note that by this interpretation rule, downward entailment is still satisfied. Because If $B \subseteq A$, then $f^a_w(\neg A) \leq f^a_w(\neg B)$, it follows that if a doubts that A and $B \subseteq A$, a also doubts that B.

Given the way that I have interpreted evidential attitude verbs and *doubt*, I have to assume that many propositions are believed. Too many, perhaps. Wouldn't it be easier and more appropriate to use probability instead of plausibility? Let P^a_w be the probability function that represents the belief state of a in w, let r be a contextually given real number in [0,1], and let α be any evidential attitude. Then it seems more appropriate to analyse the different attitude verbs in the following way:

$$[[\alpha(a, A)]]^w = 1 \quad \text{iff} \quad P^a_w(A) > r$$

$$[[\mathit{doubt}(a, A)]]^w = 1 \quad \text{iff} \quad P^a_w(A) < r$$

It is easy to see that, in the case of evidential attitudes, this analysis also accounts for closure under implication and thus for simplification. If we assume that a believes that A iff $P^a_w(A) > s$, where $0 < s \leq r$, we can also account for the belief inference (B). As regards *doubt that*, the analysis accounts for all of the principles that we want it to: addition, downward entailment, and negative simplification. This all seems pretty good; but there is a problem. Accounting for belief and evidential predicates by probabilistic means gives rise to the prediction that the relevant attitude is not closed under conjunction unless the relevant number is 1.[16] I think that this is undesirable not only for belief, but also for evidential predicates. This problem does not arise for the interpretation of *doubt that*, so there doesn't seem to be any good reason for not interpreting this predicate as suggested. Moreover, the way that we originally interpreted *doubt that* assumes that the set of believed propositions is very large. Still, it might be desirable to have a non-probabilistic account of all attitudes. Fortunately, it's not difficult to save the qualitative analysis by weakening the set $K(a, w)$. Let us follow Kratzer (1981) and use *stereotypical backgrounds* to do so. A stereotypical background is a set of worlds representing what is *normally* the case. Let us take $N(a, w)$ to represent that what a in w thinks is normally the case. Given a set of propositions, S, that potentially determines similarity, we can derive an ordinal function n^a_w by S and $K(a, w) \cap N(a, w)$, as we earlier derived the ordinal function k^a_w by S and $K(a, w)$. The belief function fn^a_w is defined via the Shackle identity in the normal way. Now we can redefine *doubt that* as follows:

$$[[\mathit{doubt}(a, A)]]^w = 1 \quad \text{iff} \quad fn^a_w(\neg A) \text{ is high}$$

Contrary to the original interpretation rule, this one does not predict that from *a doubts that A* we can infer that a believes that $\neg A$. Of course, we might have done the same by

[16] cf. section 5.4.

assuming a new accessibility relation; arguably, that is just what we did. But perhaps the new accessibility relation that we used can be used for more than just helping to analyse *doubt that*.

6.4 Desire

According to the pragmatic conception of attitudes defended in chapter 1, we can say that John desires A iff, when John's beliefs are true, John behaves in such a way that he tends to bring about that the actual world is an A-world. This puts certain constraints on how desire attributions could be analysed. Yet it still, I think, leaves open a number of alternative analyses. In the present section I will be discussing and comparing some of these alternatives.

6.4.1 A Hintikka-style analysis

In this section, we will look mainly at patterns of inference, just as we did above in analysing verbs of desire. However, we will also be taking a look at presuppositional and anaphoric relationships.

On the most straightforward account of desires, we can assume that for each agent there is a primitive accessibility relation for *desire*, Bul_j, just as there is a primitive accessibility relation for *belief*. Some have argued, however, that in contrast to the set of possible worlds for *believe that*, for desire it should not be thought of as a primitive, but should be defined in terms of the propositions desired. Because it is reasonable to assume that the propositions one desires, in contrast to those that one believes, might be mutually inconsistent, it seems like a good idea to follow van Fraasseb (1973) and Kratzer (1981) and determine an ordering relation on worlds by looking at the number of desirable propositions that the worlds make true. Thus, let $G(j, w)$ be the set of propositions that John finds desirable in w. Then we say that u is at least as desirable as v with respect to $G(j, w), u \leq_{G(j,w)} v$, iff $\{A \in G(j, w)|\ v \in A\} \subseteq \{A \in G(j, w)|\ u \in A\}$.[17] World u can now be said to be strictly more desirable than v with respect to $G(j, w), u <_{G(j,w)} v$, iff $u \leq_{G(j,w)} v$, but not $v \leq_{G(j,w)} u$. On the basis of this ordering relation we can define a function, $Bul(j, w, X)$, that gives us the set of most desirable worlds in X with respect to the ordering relation determined by $G(j, w)$:

$$Bul(j, w, X) \quad \stackrel{def}{=} \quad \{w' \in X|\ \neg\exists w'' \in X :\ w'' <_{G(j,w)} w'\}^{18}$$

[17] In this way, $\leq_{G(j,w)}$ determines a partial ordering, but not a total one. Not all worlds have to be connected with each other.

[18] One might wonder whether $Bul(j, w, X)$ should be introspective. This should be so if desires were introspective, but, as it happens, they are not: My boss wants another cigarette, but he wishes that he didn't want one.

On the basis of this function, we can now say that John desires A in w iff the set of most desirable worlds for John in w, $Bul(j, w, W)$, is a subset of A.

But this analysis immediately gives rise to a problem: it predicts that desires are closed under logical implication, but this does not seem to be the case. As noted by a number of authors,[19] if John wants A, B follows from A, and B is already believed by John, it doesn't have to be the case that he also wants B. If John hopes that his wife has survived the accident, it doesn't follow that he hopes that his wife had the accident. According to Stalnaker (1984, pp. 89-90), "the propositions one wants to be true (relative to a set of relevant possibilities) include all the consequences of any proposition one wants to be true *which distinguish between the relevant alternatives.*"[20]

What are the relevant alternatives to consider for the analysis of desire attributions? It is clear that to determine whether A is wanted or not, we have to look at a contextually-given set that contains some A-worlds and some $\neg A$-worlds. Moreover, for the analysis of *want that*, it seems that this contextually-given set is normally the set of worlds compatible with what the agent *believes*.[21] As a result, we can interpret desire attributions of the form *John wants A* in accordance with the rule given below (where $K(j, w)$ represents beliefs about the past, present, and future of John in w, and $[[A]](K(j, w))$ is the intersection of A with $K(j, w)$ if the presupposition of A is entailed by $K(j, w)$, and $\emptyset$ otherwise) and presuppose that A is true in some but not all worlds of $K(j, w)$:

$$[[Desire(j, A)]]^w = 1 \quad \text{iff} \quad Bul(j, w, K(j, w)) \subseteq [[A]](K(j, w))$$

I have argued above that desires might be mutually inconsistent. According to the interpretation rule above, the set of propositions desired, i.e. $G(j, w)$, might be mutually inconsistent; however, if A and B are mutually inconsistent, we do not predict that one can desire both A and B. The reason is that conjunction introduction is predicted to be valid. The possibility that John wants to be with his wife and that he also wants to be with his mistress, but (for obvious reasons) that he doesn't want to be with both,[22] suggests that the desires that one has need not be consistent, and thus that for the analysis of desire attributions we should look at just the most desirable worlds consistent with what is believed.

One way to solve this problem is to interpret *desire*-attributions not as necessity statements, but rather as *possibility* statements. That is, you desire A if A is *consistent*

[19] Van der Sandt (1982, 1988), Stalnaker (1984), and Heim (1992).

[20] Compare this with Dretske's reply to the sceptic in section 6.2.3.

[21] Normally, because (i) in some *want* attributions the context of interpretation for the embedded clause needs to be a *superset* of the belief state, as in Heim's (1992) example, *(John hired a baby-sitter because) he wants to go to the movie tonight*; and (ii) sometimes the context of interpretation should be a *subset* of the belief state, as in desire attributions conditionally dependent on other desire attributions, such as *John's father hopes that his son has never smoked before and hopes that he has just started smoking.* See Geurts (1995, 1998) for more on this.

[22] Some might add *at the same time or in the same place.*

with the most desirable belief worlds, as in the following rule:

$$[[Desire(j, A)]]^w = 1 \quad \text{iff} \quad Bul(j, w, K(j, w)) \cap [[A]](K(j, w)) \neq \emptyset$$

Although this rule makes rational desires closed under logical consequence, it doesn't require desires to be mutually consistent with each other, just as the wife and mistress problem discussed above appears to indicate. While this rule helps to get rid of the mutual inconsistency problem, it's not a very attractive way to analyse desire attributions. This is because it doesn't seem compatible with the pragmatic analysis of desire, according to which one desires A if one tends to bring it about that A, given that one's beliefs are true.

6.4.2 Desire as *ceteris paribus* preference

On the above analyses of desire attributions, desires are closed under implication. To account for some of the problems that such an analysis gives rise to, we have assumed that we should consider implication only with respect to the relevant alternatives. Another problem was solved by assuming that *desire* behaves like a possibility rather than a necessity operator. An alternative way to solve these problems is simply to assume that desires are *not* closed under logical implication in the first place, and to base the analysis of desire attributions more directly on preference order.

Indeed, this is what Heim (1992) argues for. She proposes that an attribution like *John wants A* is true iff John *prefers A* above $\neg A$. In this way, she gets rid of the closure condition on rational desires. The simplest possible analysis of this form would demand that preferring A above $\neg A$ means that *all* A-worlds consistent with what one believes are better than all $\neg A$-belief worlds. This would give rise to a very strong notion of desire. To weaken it, Heim assumes a *ceteris paribus* analysis of preference: A is preferred to B if for every situation compatible with what is believed, the closest world in which A but not B is true is preferred to the most similar world in which B but not A is true.[23] If we assume that f is a similarity function as defined in chapter 5 and that in w John prefers proposition X to proposition Y, $X \leq_{j,w} Y$, iff $\forall w' \in X : \ \forall w'' \in Y : \ w' \leq_{G(j,w)} w''$ & $(Y = \emptyset \Rightarrow X \leq Y)$, then Heim's interpretation rule for *want that* should go as follows:

$$[[Want(j, A)]]^w = 1 \quad \text{iff} \quad \forall w' \in K(j, w) : \ f_{w'}([[A]](K(j, w))) \leq_{j,w} f_{w'}([[\neg A]](K(j, w)))$$

If the analysis of *ceteris paribus* preference is preferable to the yes-or-no analyses of preference assumed above, then Heim's interpretation rule for desire attributions may be preferable to the above analysis of desire attributions in exactly the same sense. The advantages are that the most desirable worlds in a set are not the only ones that count and that rational desires are not predicted to be closed under logical implication.

[23] For a defense of this *ceteris paribus* analysis of preference, see Von Wright (1963) and especially Hansson (1989).

The analysis of preference implicit in Heim (1992) verifies the principle that if A is at least as desirable as B, A is also at least as desirable as $A \vee B$, which in turn is at least as desirable as B. Yet it still doesn't verify the stronger principle that if A is strictly preferred to B, and A and B are both compatible with what is believed, A is strictly preferred to $A \vee B$ which in turn is also strictly preferred to B. This principle comes out valid if we have a logic that gives $A \vee B$ a preference value somewhere *in between* the preference values of A and B. That this is needed is suggested by the following example due to Rescher (1967).

Suppose we have four relevant worlds, $\{w_1, w_2, w_3, w_4\}$, where the propositions A and B differ in truth value such that A is true in w_1 and w_2 and false in the other worlds, whereas the opposite is true for B. Suppose now that the ordering relation between possible worlds is such that w_1 is strongly preferred to w_4, which is somewhat preferred to w_2, which in turn is strongly preferred to w_3. Suppose now that except for A and B, w_1 is closest to w_3 and w_2 closest to w_4. In this situation, the *ceteris paribus* preference analysis would predict counterintuitively that A is not preferred to B and thus that A is not desired.

To clarify this with an example, let us consider the preference ordering of a German general who wants to know whether he should attack France via Belgium, A, or directly via the German-French border, B. The worlds w_1 and w_3 are very close to each other because in those worlds the French expect a German attack only directly via the German-French border. In worlds w_2 and w_4, on the other hand, the French are well prepared for a German attack both via Belgium and via the German-French border. If A is true in w_1 and w_2 and B in w_3 and w_4, clearly w_1 is strongly preferred to w_2 and w_4 to w_3. Obviously, w_1 is strongly preferred to w_3: w_1 means victory and w_3 means defeat, because the French army is assumed to be as good as the German army. It also seems reasonable to assume that if the French are prepared for an attack at both points, it is better to attack directly via the German-French border in order to reduce transport problems. So, w_4 looks a bit better to the German general than w_2. But although there is a B-world, w_4, that is strictly preferred to an A-world, w_2, the German general is advised to attack the French via Belgium and have the chance of an easy victory in battle. But according to the *ceteris paribus* analysis of preference, we should not advise the general to invade via Belgium.

How can we get rid of this problem? The answer is simple: By using a finer-grained preference logic. The most suitable logic for our purposes seems to be (a variant of) Jeffrey's (1965) preference theory, to which I will now turn.

6.4.3 Desire as quantitative preference

What is nice from our point of view is that Jeffrey's theory of preference, in contrast to some other quantitative preference logics, is compatible with the Boolean analysis of connectives common in semantics. Let us assume that $P_{j,w}$ is the probability function that assigns to each world its probability according to j in w, and that $d_{j,w}$ is a function that assigns to each possible world a real number, which indicates its desirability according to j in w. The

probability that j assigns to A in w, $P_{j,w}(A)$, is simply the sum of the probabilities of the cases (worlds) in which it is true, $P_{j,w}(A) = \sum_{v \in A} P_{j,w}(v)$. The desirability of a proposition A for j in w, $d_{j,w}(A)$, is a weighted average of the desirabilities of the worlds in which it is true, where the weights are proportional to the probabilities of the worlds,

$$d_{j,w}(A) \quad = \quad \frac{\sum_{v \in A} P_{j,w}(v) \times d_{j,w}(v)}{\sum_{v \in A} P_{j,w}(v)} \quad = \quad \frac{1}{P_{j,w}(A)} \times \sum_{v \in A} P_{j,w}(v) \times d_{j,w}(v).^{24}$$

Given Jeffrey's preference theory, the simplest way to proceed would be to say that a desire attribution *John desires that A* is true if the desirability for John of the embedded clause is greater than the desirability of a tautology:

$$\begin{array}{lll} [[Desire(j, A)]]^w = 1 & \text{iff} & \frac{1}{P_{j,w}(A)} \times \sum_{v \in A} P_{j,w}(v) \times d_{j,w}(v) > \sum_{v \in W} P_{j,w}(v) \times d_{j,w}(v) \\ & \text{iff} & d_{j,w}(A) > d_{j,w}(\top) \end{array}$$

This can easily be seen to have the following benefits: (i) it doesn't predict that desire will be closed under logical consequence; (ii) it doesn't preserve the validity of conjunction introduction; (iii) it predicts that if $Desire(j, A)$ is true and John prefers B to A, then $Desire(j, B)$ is also true; and (iv) it can account for Rescher's problem.[25] In contrast to Heim's analysis of buletic predicates, it doesn't make use of the *ceteris paribus* condition, but in this case such a condition is not needed to get a very weak system.

Let's consider again our model with four worlds, where w_1 and w_3 on the one hand, and w_2 and w_4 on the other, are most similar to each other. Let us also assume that $A = \{w_1, w_2\}$, $B = \neg A = \{w_3, w_4\}$, and all four worlds are equally likely to be true. In this case, the *ceteris paribus* analysis of preference demands that for A to be desired, w_1 and w_2 must be preferred to w_3 and w_4, respectively. Jeffrey's preference theory, in contrast, demands only – assuming a cardinal valuation to the four worlds – that the average valuation of w_1 and w_2 must be higher than the average valuation of w_3 and w_4. As this example illustrates, the quantitative approach *weakens* Heim's qualitative approach. On the quantitative approach, we don't compare possible worlds that are most similar to each other, but instead compare whole information states. I am not sure whether this weakening is in general superior to Heim's strong notion of preference, but as the above example of the German general shows, it seems to be superior in at least some cases.

6.4.4 A conditional analysis of desires

So far we have discussed four kinds of analyses of desire attributions. The first two were based on a classical all-or-nothing analysis of preference, the second on a *ceteris paribus*

[24] For simplicity, I am assuming in this formulation that there are only finitely many possible worlds. If there are infinitely many worlds in which a certain proposition is true, every world in this set has probability 0. It is important, however, that Jeffrey's theory does not require these assumptions. Desirability can also be defined for continuous probability functions, in which case we need intervals and integrals.

[25] Rescher's (1967) logic of preference can also handle those problems. But this is no big surprise since Rescher's logic is only a special case of Jeffrey's system. For Rescher all possible worlds have equal probability. It is clear that this makes Rescher's logic less suitable for decisions under uncertainty.

analysis of preference, and the third on a quantitative notion of preference. In this section I will discuss yet another analysis of preference.

Asher (1987) observes that desire attributions normally obey disjunction elimination, and Zimmermann (ms.) observes that indefinites in the scope of verbs of desire are normally interpreted 'arbitrarily'. Thus, we can normally infer (116b) from (116a), and interpret (117a) as something like (117b):

(116) a. Alexis hopes that she will have chicken or fish for dinner.

b. So she hopes that she will have chicken for dinner.

(117) a. John wants to catch a fish.

b. John wants to catch an arbitrary fish; any fish will do.

These facts are surprising for any of the above proposals. They can be accounted for properly for, however, if we assume that desire attributions are understood as implicit conditionals. Thus, *John wants that A* means something like 'If A is the case, John will be satisfied'. Disjunction elimination now follows immediately; unfortunately, the more general *downward entailment* is also predicted to hold. That is, if John wants A, and B entails A, it would follow that John wants B too. But this is obviously wrong: I want to have a holiday this summer, but not a holiday and bad weather. Yet the conditional interpretation of desire attributions can still be rescued if this conditional is treated not as an indicative conditional but as a *subjunctive* conditional instead. To make sense of this, we can assume that $K(j, w)$ represents the possible ways the world might be *at this moment* according to John in w, rather than the set of futures consistent with what John believes in w, as we have been assuming.[26] Thus, if we want to look at the future, we have to use the more general revision rule. I will assume that if somebody wants A, they have a desire about the future and so do not believe A yet. Desire attributions can now be analysed in terms of revision as follows:

$$[[Desire(j, A)]]^w = 1 \quad \text{iff} \quad C''_{K(j,w)}(A) \subseteq Bul(j, w, W)^{27}$$

Thus, *John wants A* is true in w iff $K(j, w)$ revised by A is a subset of the set of John's absolute favorites among (what he considers to be) the possible futures. Note that according to the above rule, neither upward nor downward entailment is valid. Moreover, disjunction elimination is allowed, but only if the complements of both disjuncts are equally

[26] I have in mind here the framework of branching time as developed by, among others, Thomason & Gupta (1980). The details of this analysis need not worry us here.

[27] Where $Bul(j, w, W)$ is defined as in section 6.4.1. The form of this interpretation rule was actually proposed by Price (1989) in his defense of the Desire-as-Belief thesis. An alternative formulation would be to use imaging, defined in terms of a fixed selection function f. In this case, as we saw in chapter 5, we would not expect that it always holds that if A is consistent with $K, C_K(A) = K \cap A$.

strongly entrenched.[28] This seems exactly what we need. Normally, disjunction elimination is valid and indefinites receive an arbitrary interpretation; but this is not always the case:

(118) John wants a beer, but not a warm one.

6.4.5 Buletic ordering

Still, a counterexample like (118) to the arbitrary interpretation of the indefinite has intuitively nothing to do with *epistemic* entrenchment. This suggests that the ordering relation by which we determine the relevant change function should not be induced by epistemic entrenchment but by *desirability* instead. What we could do is to take up the consistency interpretation of desire attributions of section 6.4.1, but consider not whether there are A-worlds among the most desirable belief-worlds, but rather whether *the best* A-worlds are among the most desirable belief worlds. This suggests that we should use the following interpretation rule:

$$[[Desire(j, A)]]^w = 1 \quad \text{iff} \quad Bul(j, w, [[A]](K(j, w))) \subseteq Bul(j, w, K(j, w))$$

This interpretation rule has a number of desirable consequences. First, it predicts that disjunction elimination and the arbitrary interpretation of indefinites used in desire attributions are not valid. From *John wants that A or B*, I can conclude only that John also wants A, if A is at least as desirable for John as B is. Similarly, from *John wants an apple* I can conclude only that John wants a green apple, if eating green apples is at least as desirable for John as eating apples of any other color. And this is confirmed by (118). Second, in contrast to the conditional analysis given in the previous section, this analysis doesn't need to resort to revision in order to avoid the prediction that if B entails A, desiring A entails desiring B. This is because we look only at the best A-worlds compatible with what is believed. Third, it can account for the unacceptability of sequences like (119):

(119) John wants a cool beer, but he doesn't want a beer.

On this approach, the reason why (119) is unacceptable is that desires are closed under logical implication. It can easily be verified that the above interpretation rule make the sentence *John wants A* true in w for any A compatible with what is believed iff $[[A]](K(j, w) \cap Bul(j, w, K(j, w)) \neq \emptyset$. From this it immediately follows that if $A \subseteq B$, it is also the case that $[[B]](K(j, w)) \cap Bul(j, w, K(j, w)) \neq \emptyset$, and thus that John wants B too.

[28] If the revision function C'' obeys $(R^*1) - (R^*4)$, $C''_K(A \vee B) = C''_K(A) \cup C''_K(B)$, if $\neg A$ and $\neg B$ are equally strongly entrenched in K.

6.4.6 Combining belief revision and desirability

According to the above analysis, all counterexamples to disjunction elimination arise because some disjuncts are *strictly preferred* to other disjuncts. Although this does appear to be the reason behind many such counterexamples; I don't believe that it is the reason behind all of them.[29] Consider (116a)-(116b) again, repeated as (120a)-(120b):

(120) a. Alexis hopes that she will have chicken or fish for dinner.

b. So she hopes that she will have chicken for dinner.

Consider now the case where Alexis thinks that there is a tiny chance of having chicken, A, and a good chance of having fish, B. She prefers both to anything else she considers possible, but has no preference for one above the other, i.e., in w it holds that $A \approx_{a,w} B$. The above analysis, just like the quantitative analysis discussed above, would then predict that disjunction elimination is allowed. However, it seems that in such circumstances it is fine to assert (120a) but not (120b).

Perhaps the most obvious way to account for this is to make use of the quantitative framework. By using Jeffrey's theory of preference, we might say that instead of looking at the *desirability* of a proposition, we should look instead at its *expected value* or *utility*. Where the desirability of a proposition is the *weighted* average of the desirabilities of the worlds in which it is true, and thus does not increase in the case the probability of the proposition increases, the expected value of a proposition increases in the case the probability increases. The expected value of A for John in w, $EV_{j,w}(A)$ is defined as follows:

$$EV_{j,w}(A) \quad = \quad \sum_{v \in A} P_{j,w}(v) \times d_{j,w}(v).$$

Then we might say that each desire attribution is interpreted with respect to a set of alternatives, C; and that John desires A if the expected value of A is at least as high as the expected value of any of its alternatives:

$$[[Desire_C(j, A)]]^w = 1 \quad \text{iff} \quad \forall B \in C : \ EV_{j,w}(A) \geq EV_{j,w}(B)$$

For our example this means that having chicken for dinner has a lower expected utility than having chicken or fish, because of the high probability of having fish.

Another way to account for the above problem within a quantitative framework is to make use of the *revision* of probability functions. We can do this by making use of *Popper functions*, also known as *extended probability functions*.[30] Popper functions are probability functions that take conditional probabilities as basic. So for a Popper function Pr, unlike for standard probability functions, $Pr(A/B)$ is also defined if $Pr(B) = 0$. As a result, such

[29] I am indebted to Ede Zimmermann (personal communication) for this.

[30] See section 5.4.

a function contains the extra information about what would happen under revision. Harper (1976a) shows that if we limit ourselves to probability 1, the minimal revision modelled by Popper functions satisfies the standard revision rules $(R^*1) - (R^*4)$. Let us now say that $Pr_{j,w}(v/A)$ gives us the probability John assigns to v in w under the revision of A. In that case we can define the *desirability* of A, $d_{j,w}(A)$, with respect to probability function $Pr_{j,w}$ and desirability function $d_{j,w}$ as follows:

$$d_{j,w}(A) \quad = \quad \sum_v Pr_{j,w}(v/A) \times d_{j,w}(v).$$

Observe that this is similar to the *expected value* of A. Now that we have made use of revision, however, we can say that one desires A if the expected value of A is greater than the expected value of doing nothing, $d_{j,w}(\top)$.

The above solution to Zimmermann's problem also has its qualitative variants. Note that on Heim's (1992) analysis, desirability and update already play separate roles. If in Heim's interpretation rule for desire attributions we now exchange the update function $[[.]]$ by the revision function C'', the inference from (120a) to (120b) in the above situation does not go through. This is because if $\neg A$ is more strongly entrenched in $K(a, w)$ than $\neg B$, then $C''_{K(a,w)}(A \vee B)$ will be incompatible with A.

The conditional analysis discussed in section 6.4.4 also seemed to make the right predictions. The only problem was that it only considered *epistemic* and not *buletic* entrenchment. After discussion of this problem, we analysed desire attributions in the following section by making use of only *buletic* entrenchment, but noted that this was not adequate. A natural proposal then, is simply to combine the two approaches. One way of doing this is to analyse desire attributions as follows:

$$[[Desire(j, A)]]^w = 1 \quad \text{iff} \quad Bul(j, w, C''_{K(j,w)}(A)) \subseteq Bul(j, w, K^*(j, w))$$

In this interpretation rule, $K(j, w)$ denotes the set of ways the world might be *at this moment* according to John in w, whereas $K^*(j, w)$ denotes the set of futures consistent with what John believes. Note that with this interpretation rule we can account for Zimmermann's problem. Remember that the problem was that both propositions of the disjunction $A \vee B$ were equally desirable; but that one, B, was more likely to be or to become true than the other. The desire is about the future, so $K(j, w)$ will be inconsistent with both. However, Alexis considers it more likely that she will have fish, B, than chicken, A; and thus $C''_{K(j,w)}(A \vee B)$ will contain only B-worlds. As a result, we predict that in this situation, (120a) is true but (120b) false, just as we want.

What might be worrying about our two approaches above is that by making use of belief revision, we no longer predict that desires are closed under implication. The reason is that the revision function C'' does not obey the following monotonicity condition: $A \subseteq B \Rightarrow C''_K(A) \subseteq C''_K(B)$. As a result, if normal beer-drinker John considers it equally likely that he will get a warm or a cool beer, then sentence (119), repeated here as (121), will be predicted to be true:

(121) John wants a cool beer, but he doesn't want a beer.

As we have seen above, however, this sentence is infelicitous. Earlier on, we accounted for this by assuming that desires are closed under logical implication. Now I want to suggest that we don't have to make this assumption in order to explain the infelicity of (121), once we make use of *modal subordination*, as described in chapters 3 and 4.

According to this explanation we have to assume that when we make a sequence of desire attributions, only the first one is interpreted with respect to the belief state of the agent, while the second one is interpreted with respect to an information state in which the first desire is fulfilled. On this assumption, the second attribution of a sequence of the form $Desire(j, A) \wedge \neg Desire(j, A \vee B)$ would be trivially false if A is inconsistent with $K(j, w)$. Once the assertion of a sentence is trivially false, it is *inappropriate* to make it. Thus, we don't have to assume that desires are closed under logical implication in order to explain why (121) is odd.

Although we have been discussing verbs of desire in general, in most of this discussion we have clearly had our sights on the verbs *want that* and *hope that*. This was especially the case with my assumption that only possibilities consistent with what the agent believes are relevant in the analysis. Once we make use of belief revision or contraction, however, it also becomes possible to account for *factual* and *counterfactual* desire attributions like *John is glad that A* and *John wishes that A*, respectively. For the analysis of counterfactual desire attributions, for instance, the only thing we need to do is to exchange the set $K^*(j, w)$ for W; and for a factual desire attribution like *John is glad that A*, we would (also) need to exchange the set $K(j, w)$, where A is already assumed, for the contracted belief state, where A is given up.[31]

I haven't said anything about how the different analyses described above could account for *anaphoric* dependencies across desire attributions. At this point I only want to note that once we make use of belief revision for the analysis of desire attributions, an account of such dependencies becomes quite straightforward. In this context, consider the following examples:

(122) Sue wants to marry *a Swede*, and she wants a child from *him*.

(123) John wants to catch *a fish*, and he wants to eat *it* afterwards.

On their most natural interpretations, the indefinites in the first clauses of these sentences are not used by the speaker in a specific sense. They are intended to refer neither to a specific Swede or fish nor to a specific 'belief object' that the agent has. Thus, according to the theory of anaphora I defended in chapter 2, the pronouns occurring in the second clauses can be used only as *descriptive pronouns*. But in order for a descriptive pronoun to be used appropriately, the relevant property introduced by the antecedent indefinite

[31] Harper (1977) was the first, as far as I know, to define the contraction of K by A as $K \cup C''_K(\neg A)$.

must be presupposed to satisfy the *uniqueness* constraint in each possibility of the relevant context of interpretation. The crucial point here is that it is very natural to assume that this uniqueness constraint is satisfied once we analyse desire attributions basically as subjunctive conditionals: we go to the closest worlds where John, for instance, catches a fish; and worlds where he catches only one fish are closer to the belief worlds where he has not yet caught one than worlds where he catches more than one.

6.4.7 Intention and action

Until now I have assumed that all verbs of desire should be analysed in the same way – in other words, that the emotive cognitive attitude *hope* should be analysed in the same way as a pro-attitude like *intend*. Intuitively, however, there are at least two differences between *intend* and *hope*: (i) whereas what you intend has typically something to do with your own activities, hopes are not so closely related to the actions of the agent himself; and (ii) whereas *intend* is necessarily future-oriented, *hope* need not be, as in *I hope he survived the operation*. The verb *intend* normally takes as complement *to*-infinitives that seem to designate abilities; but, as Portner (1997) observes, the verb *hope* takes both *to* and *that*-clauses as complements. I don't want to suggest that the two verbs should be analysed in a completely different way. Yet it might be the case that we have two quite different concepts of desire – one related to futures that the agent can influence himself and one related to circumstances he cannot influence, – and that these two concepts are typically expressed by two different verbs: namely *intend* and *hope*, respectively.

One option to 'explain' this all is to say that the truth conditions for these constructions are identical and should be analysed as before, but that appropriateness conditions for asserting such sentences differ. For *hope* it should be the case that both the embedded clause and its negation should be consistent with what the agent believes about the present and with the global context, but for *intend* this need not be the case.

Perhaps the intuitive difference between the two concepts can be accounted for in this way, but it might, in fact, be necessary for us to take the notion of *action* more seriously than we have been doing so far. The obvious suggestion would be to follow causal decision theory and make use not of conditionalisation but of *imaging*.[32,33] For instance, we could say that you intend A, or intend to make A true, if the *utility* of A is higher than doing nothing or higher than any other relevant alternative action/proposition, where the utility of A, $u(A)$ is defined as $\sum_w P_A(w) \times d(w)$. Alternatively, we might use the conditional

[32] As discussed in section 5.6.4.

[33] It is tempting at this point to analyse actions not simply in terms of imaging, but in terms of *dynamic logic*, making use of *programs*. But you might think that analysing actions in terms of imaging just involves assuming that all actions are *atomic* programs: both simply denote functions from worlds to sets of worlds, and the function that they denote is determined by the model. However, once action-denoting sentences become complex, the two analyses will typically give rise to different results. In particular, while change by imaging is not *monotonic*, change by applying programs is. Note that by analysing actions in terms of dynamic logic, we might get rid of a problem for *intend* that is analogous to (121).

analysis for intention and say that John intends A in w iff doing A ensures that John fulfills his goals:

$$[[intend(j, A)]]^w = 1 \quad \text{iff} \quad C_{K(j,w)}(A) \subseteq Bul(j, w, K^*(j, w))$$

Notice that neither of the two analyses predicts that intentions are closed under the believed consequences or side effects of these intentions. Bratman (1987) and Cohen & Levesque (1990) argue that intentions should, indeed, not be closed under believed consequences. Consider Susan, who has a toothache. She intends to get rid of the toothache by getting her tooth filled. She believes, however, that getting her tooth filled will cause her much pain, because she is not well informed about anaesthetics. Still it seems reasonable to assume that she does not have the intention to be in pain.

Although our analyses predict that intentions are not closed under the consequences believed to follow from them, they do predict that for any A and B that are believed to causally entail each other, i.e. $C_{K(j,w)}(A) = C_{K(j,w)}(B)$, it follows that by intending one, you automatically also intend the other. However, it seems that even this is too strong a prediction: Suppose John intends to become rich, A, and believes that the only way to do so is to work very hard, B. Thus, John believes that $A \rightsquigarrow B$ is true, where $\rightsquigarrow$ is our non-backtracking counterfactual connective. But John also has a lot of faith in himself, and believes that if he worked hard, he would also become rich. So he also believes $B \rightsquigarrow A$. In other words, the condition $C_{K(j,w)}(A) = C_{K(j,w)}(B)$ is satisfied. Still, in at least one sense of the word, I can imagine John intending to become rich but not intending to work very hard: if John found out that he could become rich without working hard, he *would* go for that option.

What this argument suggests is that you can intend something only if your desire or goal to perform the intended action is relatively immune, or *stable*, *under belief revision*. According to Bratman (1987), it is this stability of intentions that make them so useful for agents: we don't have to deliberate at each moment whether or not to perform a certain action.

If intention is an attitude that is relatively stable under belief revision, it shares a lot with another attitude, the attitude of *knowledge*, as discussed in section 2.3 of this chapter. It even seems plausible to analyse intention partly in terms of knowledge. I want to suggest tentatively that *John intends A* iff (i) John desires to do A, and (ii) almost no amount of further information would change that desire. A crude way to implement this is to say that doing A satisfies an agent's desires with respect not only to his *belief* alternatives, but also to his *epistemic* alternatives:

$$[[intend(j, A)]]^w = 1 \quad \text{iff} \quad C_{EPI(j,w)}(A) \subseteq Bul(j, w, EPI^*(j, w))$$

Notice that we now predict that even if John believes that A and B are causal consequences of each other, he can still intend the one without intending the other. The stronger condition that needs to be fulfilled now is that John must *know* that the two are causal

consequences of each other: $C_{EPI(j,w)}(A) = C_{EPI(j,w)}(B)$ – a condition that in our above example is probably not fulfilled.

The precise way to implement intention is not as important as the main idea behind it. This is that *intention* is a robust kind of *desire*, just as *knowledge* is a robust kind of *belief*.

Appendix A

Two-dimensional counterpart theory

In this appendix I will formalize most of what I have argued for in Chapter 1 in the framework of quantified modal logic. But the formalization is unusual in a number of ways. First, the logic will be a *free* logic; singular terms may fail to denote. Second, following Stalnaker (1977), I assume the only variable-binding operator in our language is the *abstraction operator* that turns sentences into complex predicates. As a result, we (i) can assume that the quantifier directly applies to a predicate, and (ii) can define complex singular expressions (iota terms). Third, our analysis is a *token analysis*; the interpretation function is not defined on types of expressions, but rather on tokens of expressions. For simplicity I will assume that the token analysis is only relevant for the analysis of indexicals. Fourth, the framework is *two-dimensional*; I assume that every token is interpreted with respect to *two* worlds, a context-world, and an index world. The two-dimensionality is of course used to account for diagonalisation. Fifth, to account for aboutness, the logic will be partial; atomic formulae need not be true or false with respect to all indices. Finally, my analysis is a *counterpart theory*; I assume the domains of different worlds are disjoint, and that in order to determine the truth value of modal statements or belief attributions *about* particular individuals, we have to look at the/a counterpart of this particular individual in other relevant worlds.[1]

Syntax

I will now define the utterance language **L** whose expressions are sets, like in Cresswell (1973). The expressions of **L** are either symbols, terms, or complex expressions. The language **L** has the following symbols:

(i) basic symbols: $\neg$, $\wedge$, $\forall$,), (, $\hat{}$, $\Box$, Bel, ι, $=$;
(ii) a denumerable set of individual variables: $VAR_L = \{x, y, ...\}$;
(iii) individual constants: $CONST_L = \{a, b, ...\}$;

[1] See Aloni (2001) again for another formalization of counterpart theory.

(iv) the set of demonstratives: $DEM_L = \{I, you, here, now, ...\}$;
(v) for every $n \geq 0$, a denumerable set of primitive n-place predicates.

The language **L** is defined by the following definition of the terms, and complex expressions of **L**:

The set of *terms* of **L**, $TERM_L$, is equal to $VAR_L \cup CONST_L \cup DEM_L \cup CSTERMS_L$, where $CSTERMS_L$ is the set of complex singular terms of **L** to be defined below.

The set of *complex expressions* of **L** is defined as follows:

(a) Sentences

(i) If $t_1...t_n$ are terms and P an n-place predicate, then $Pt_1...t_n$ is a sentence.
(ii) If t_1 and t_2 are terms, then $t_1 = t_2$ is a sentence.
(iii) If A is a sentence, then $\neg A$ is a sentence.
(iv) If A and B are sentences, then $A \wedge B$ is a sentence.
(v) If P is a one-place predicate, then $\forall P$ is a sentence.
(vi) If A is a sentence, and t is a term, then $\Box A$ and $Bel(t, A)$ are sentences

(b) Complex predicates:

(i) If A is a sentence and x a variable, then $\hat{x}A$ is a one-place predicate.
(ii) If P is an n-place predicate, then $\dagger P$ and $@P$ are also n-place predicates.

(c) Complex singular terms:

(i) If P is a one-place predicate, then ιP is a complex singular term.
(ii) If t is a term, then $\dagger t$ and $@t$ are complex singular terms.

There are no other complex expressions.

The formulae $\exists P$ and $\Diamond A$ will be the abbreviations of $\neg\forall\neg P$ and $\neg\Box\neg A$, respectively.

By a token of an expression a of **L** we simply mean a sequence whose first member is a member of the first member of a, whose second member is a member of the second member of a, and so on. I will for simplicity do as if an underlined expression denotes a particular token of this type of expression. If b is a primitive symbol, I will write $\underline{b}$ for a token of b, i.e. an element of it. Officially, a token of a complex expression like $Pt_1, ...t_n$ is of the form $\langle \underline{P}, t_1, ..., t_n \rangle$, but I will simply write it as $\underline{Pt_1, ..., t_n}$. For simplicity again I will sometimes misuse the language saying that also for a complex expression b it holds that $\underline{b}$ is an element of b. If A is a complex expression, a token of A consists of more than one token. In particular, it might be the case that a token of one complex expression contains two or

more tokens of the same primitive expression that should become different denotations. As when Michael Corleone orders *You take care of Dani Sciandri, you deal with the De Vito brothers, and you, ..., you make coffee.* Thus, it is the tokens of the smallest symbols that count, not those of the complex expressions. Now we give a semantics for the utterance language **L**.[2]

Semantics

Pointed models are nine-tuples $\langle W, w_0, D', T, *, R, C, \{K_a\}_{a\in A}, I\rangle$, where W is a non-empty set of worlds, w_0 a designated element of W, representing the actual world, D' is a function from W to a non-empty set such that for any two different worlds w and w', $D'(w) \cap D'(w') = \emptyset$. I will also denote $\bigcup\{D'(\mathrm{w})|w \in W\}$ by D'. T is a set of tokens of **L**. I will assume that the same token of **L** can occur in many elements of W. The union of D' and T is denoted by D. $*$ is a special object that is not an element of D, R a binary relation on W that is reflexive, transitive, and symmetric, C a set of total functions in $[((D' \cup \{*\}) \times W) \rightarrow D' \cup \{*\}]$, the counterpart functions, the set of belief functions $\{K_a\}_{a\in A}$, where each K_a is a function in $[W \rightarrow \wp(W)]$, such that for each w and w' : $w' \in K_a(w) \rightarrow K_a(w') = K_a(w)$, mirroring that belief states are introspective,[3] and I the interpretation function. The interpretation function I meets the following conditions:

- for any two tokens $\underline{c}$ and $\underline{c'}$ of an individual constant symbol c and for each world w:

  $I(\underline{c})(w) = I(\underline{c'})(w) \in (D' \cup \{*\})$[4]

- for any two tokens $\underline{P}$ and $\underline{P'}$ of any primitive n-ary predicate symbol P and any $w, w' \in W : I_{w,w'}(\underline{P}) = I_{w,w'}(\underline{P'}) \subseteq (D(w'))^n$

If $\underline{P}$ is a token of a primitive predicate symbol, then $I_{w,w',c,g}(\underline{P}) = I_{w,w',g}(\underline{P}) = I_{w,w'}(\underline{P})$.[5]

[2] I will give a direct semantics for *tokens*, not for types. You might object, saying that it is *types* that have a semantic value, not tokens, and that all that tokens do is to single out the context relative to which the character of the semantic type has to be evaluated. I agree, but you might read my interpretation rules as doing simply two steps in one. My semantics is not so much a semantics for tokens as such, but rather a semantics for tokens *as tokens of a type of a particular language.*

[3] The subscript a should be thought of as a constant denoting an individual concept, that denotes in each world in $K_a(w)$ how the individual denoted by a in w thinks about himself.

[4] I make here the simplifying assumption that for each particular name, we refer to the same individual, on any occasion we use it. Of course, this assumption can be given up, but it would make things much more complicated.

[5] Note that also tokens of primitive predicates are interpreted with respect to two worlds. In this appendix, however, I will assume that the distinction between the worlds is irrelevant here.

All counterpart functions obey the following constraints:[6]

- $\forall w \in W, c \in C, d \in D'(w) : c_w(d) = d$
- $\forall w \in W, c \in C, d \in D' : c_w(d) \in (D'(w) \cup \{*\})$
- $\forall w \in W, c \in C : c_w(*) = *$

An assignment function is a function mapping an individual, an element of D', to tokens of variables. G is the set of all assignment functions belonging to $\mathbf{L}$. I will assume that for each individual variable symbol x, and for each $\underline{x}$, $\underline{x'} \in x$ it holds that for any element g of G: $g(\underline{x}) = g(\underline{x'})$. I will assume that $g[^x/_d]$ is an abbreviation for $\{\langle \underline{y}, g(\underline{y})\rangle |\ \underline{y} \in dom(g)\ \&\ \underline{y}$ is not a token of variable symbol $x\} \cup \{\langle \underline{x}, d\rangle |\ \underline{x}$ is a token of variable symbol $x\}$.

Tokens of individual terms will be interpreted in terms of a counterpart function and the object denoted by $[\underline{t}]^{w,w',g}$:

$$[[\underline{t}]]^{w,w',c,g} \quad = \quad c_{w'}([\underline{t}]^{w,w',g})$$

The object denoted by $[\underline{t}]^{w,w',g}$ is determined as follows:

$[\underline{t}]^{w,w',g}$ = the utterer of $\underline{t}$ in w, if there is one, and if $\underline{t}$ is a token of I
(and so on for the other demonstratives)
= $I(\underline{t})(w)$, if $\underline{t}$ is a token of a constant symbol;
= $g(\underline{t})$, if $\underline{t}$ is a token of a variable;
= $[\underline{t'}]^{w',w',g}$, if $t = \dagger t'$ for some $t' \in TERM_L$
= $[\underline{t'}]^{w,w,g}$, if $t = @t'$ for some $t' \in TERM_L$
= d, if $\underline{t} = \underline{\iota P}$ and $I_{w,w',g}(\underline{P}) = \{d\}$;
= $*$ otherwise.

The satisfaction conditions are defined as follows (neglecting the subscript for the model):[7]

(1a) $[[\underline{t_1 = t_2}]]^{w,w',c,g} = 1$ iff $[[\underline{t_1}]]^{w,w',c,g} = [[\underline{t_2}]]^{w,w',c,g}$

(1b) $[[\underline{t_1 = t_2}]]^{w,w',c,g} = 0$ iff $[[\underline{t_1}]]^{w,w',c,g} \neq [[\underline{t_2}]]^{w,w',c,g}$

[6]Note that just as different constraints on accessibility relations in standard modal logic would give rise to different logics, in counterpart theory different constraints on the counterpart functions would give rise to different logics too.

[7]I will say that $[[A]]^\alpha = c$ iff $[[A]]^\alpha \notin \{a, b\}$, if a, b, and c are the three truth values. Note that the resulting logic corresponds with Bochvar's (1939) system, very similar to the four valued logic we use for the analysis of presuppositions in chapter 4.

(2a) $[[\underline{P(t_1,...,t_n)}]]^{w,w',c,g} = 1$ iff $\langle [[\underline{t_1}]]^{w,w',c,g}, ..., [[\underline{t_n}]]^{w,w',c,g}\rangle \in I^{+}_{w,w',c,g}(\underline{P})$[8]

(2b) $[[\underline{P(t_1,...,t_n)}]]^{w,w',c,g} = 0$ iff $\langle [[\underline{t_1}]]^{w,w',c,g}, ..., [[\underline{t_n}]]^{w,w',c,g}\rangle \in I^{-}_{w,w',c,g}(\underline{P})$ [9]

(3a) $[[\underline{\neg A}]]^{w,w',c,g} = 1$ iff $[[\underline{A}]]^{w,w',c,g} = 0$

(3b) $[[\underline{\neg A}]]^{w,w',c,g} = 0$ iff $[[\underline{A}]]^{w,w',c,g} = 1$

(4a) $[[\underline{A \wedge B}]]^{w,w',c,g} = 1$ iff $[[\underline{A}]]^{w,w',c,g} = 1$ and $[[\underline{B}]]^{w,w',c,g} = 1$

(4b) $[[\underline{A \wedge B}]]^{w,w',c,g} = *$ iff $[[\underline{A}]]^{w,w',c,g} = *$ or $[[\underline{B}]]^{w,w',c,g} = *$

(5a) $[[\underline{\forall P}]]^{w,w',c,g} = 1$ iff $I_{w,w',c,g}(\underline{P}) = D(w')$

(5b) $[[\underline{\forall P}]]^{w,w',c,g} = 0$ iff $I_{w,w',c,g}(\underline{P}) \cap D(w') \neq \emptyset$

(6) $[[\underline{\dagger A}]]^{w,w',c,g} = 1/0$ iff $[[\underline{A}]]^{w',w',c,g} = 1/0$

(7) $[[\underline{@A}]]^{w,w',c,g} = 1/0$ iff $[[\underline{A}]]^{w,w,c,g} = 1/0$

(8a) $[[\underline{\Box A}]]^{w,w',c,g} = 1$ iff $\forall w'' \in R(w') : [[\underline{A}]]^{w,w'',c,g} = 1$

(8b) $[[\underline{\Box A}]]^{w,w',c,g} = 0$ iff $\exists w'' \in R(w') : [[\underline{A}]]^{w,w'',c,g} = 0$

(9a) $[[\underline{Bel(t,A)}]]^{w,w',c,g} = 1$ iff $\forall w'' \in K([[\underline{t}]]^{w,w',c,g}, w') :\ [[\underline{A}]]^{w,w'',c,g} = 1$

(9b) $[[\underline{Bel(t,A)}]]^{w,w',c,g} = 0$ iff $\exists w'' \in K([[\underline{t}]]^{w,w',c,g}, w') :\ [[\underline{A}]]^{w,w'',c,g} = 0$

If $\underline{P}$ is a token of a complex predicate of the form $\hat{x}A$ (where $A \in FORM_L$), then

$$I^{\pm}_{w,w',c,g}(\underline{P}) = \{d \in D(w') : [[\underline{A}]]^{w,w',c,g[^x/_d]} = 1/0\};$$
$$I^{\pm}_{w,w',g}(\underline{P}) = \{d \in D(w') : \exists c \in C :\ [[\underline{A}]]^{w,w',c,g[^x/_d]} = 1/0\}.$$

Although normally a sentence is interpreted with respect to the contextually given counterpart function, I will assume that *semantically* we *existentially* quantify over counterpart functions. Thus, for any token $\underline{A}$ of a formula A, the absolute notion of satisfaction is defined as follows: $\underline{A}$ is *satisfied* with respect to w, w' and g, $w, w', g \models \underline{A}$, iff

[8] The existence predicate is defined as $\hat{x}\exists\hat{y}(y = x)$.

[9] I assume that if $\underline{P}$ is a token of a primitive predicate, $I^{-}_{w,w',c,g}(\underline{P}) = (D(w'))^n - I^{+}_{w,w',c,g}(\underline{P})$. Of course, I hereby do not go completely classical, because $*$ is not an element of D.

$\exists c \in C: \ [[\underline{A}]]^{w,w',c,g} = 1$. Now we can define a notion of truth with respect to a context world w (and a model), and an absolute notion of truth (with respect to a model): $\underline{A}$ is *true in* w' *with respect to* w, $w, w' \models \underline{A}$, iff for all $g \in G : w, w', g \models \underline{A}$, and $\underline{A}$ is *absolutely true* iff $\underline{A}$ is true in w_0 with respect to w_0. Finally, we say that $\underline{A}$ is *valid*, $\models \underline{A}$, iff $\underline{A}$ is absolutely true in all models.

Discussion

Note that in our *two-dimensional* system we allow for two kinds of propositions expressed by an utterance, the *horizontal* proposition, and the *diagonal* proposition. The horizontal proposition expressed depends on the particular counterpart function that is chosen, and on the context-world. If this counterpart function is c, and the context-world is the actual world of the relevant model, w_0, then the horizontal proposition expressed by utterance $\underline{A}$ in this model is $\{w' \in W | w_0, w', c \models \underline{A}\}$. On the other hand we have what Stalnaker (1978) calls the *diagonal* proposition expressed by $\underline{A}$, the proposition that is true in w iff the horizontal proposition expressed by $\underline{A}$ in w is also true in w.

Note that according to this semantics an utterance of the form *I am now here* is *valid* or *a priori* true, absolutely true in all models, although the horizontal proposition expressed by it need not be *necessarily* true, true in all metaphysical accessible worlds of, and with respect to, the actual world of the model, in any of the models. Thus, our system shares with Kaplan's (1989) logic the feature that validities are not closed under necessitation. In other words, our system allows utterances to express *contingent* propositions, although they are still *a priori* true. Also the utterance *Deep Throat is the person in the White House who was the source of Woodward and Bernstein's Watergate information* can be said to express a contingent *a priori* truth, if it's assumed that we *actually* speak English, that is, that in all distinguished worlds of all models the actual conventions of English are obeyed.[10]

Diagonalisation is, as we have seen in the main text, also crucial to account for many puzzles that arise in *epistemic* contexts, especially for *de dicto* and *de se* belief attributions. For instance, although each of the pairs 'Hesperus' and 'Phosphorus', 'a fortnight' and 'a period of fourteen days', and 'woodchucks' and 'groundhogs' actually have the same denotation, we can make sense of the intuition that belief attributions like *John does not believe that Hesperus is Phosphorus*, *John does not believe that a fortnight is a period of fourteen days* and *John believes that no woodchuck is a groundhog* still might be true.

Note that this semantics satisfies the rigidity assumption for individual constants: $t = t' \rightarrow \Box(t = t')$, and $t = t' \rightarrow Bel(a, t = t')$, for any two elements t and t' of $CONST_L$.[11]

[10] Of course it's not the same object that is both contingent and *a priori* true. The *horizontal* proposition expressed by these sentences is contingent, but the *diagonal* proposition expressed is *a priori* true.

[11] From now on I tend to forget that tokens of expressions should be underlined, and that officially I always make use of abstraction operators.

More generally, the formula $\forall x \forall y[x = y \rightarrow \Box(x = y)]$ is valid. The most important formal distinctions between our counterpart modal logic and standard quantified modal logic are, if we ignore belief attributions, that according to this semantics the clause for quantification is world-dependent; that (using standard notation) the formula $\forall x \Box \exists y(y = x)$ can be false; that $\exists x \exists y[x \neq y \wedge \Diamond x = y]$ is satisfiable; and that the principles of existential generalization (EG) and universal instantiation (UI) are no longer valid. The reason that (EG) and (UI) are no longer valid is that singular terms do not have to refer to an object in the domain of quantification. More interesting is that the Free Logic versions of (EG) and (UI),

(FEG)	$(A(t) \wedge E(t)) \rightarrow \exists x Ax$	(E is the existence predicate)
(FUI)	$\forall x A \rightarrow (E(t) \rightarrow A(t))$,	for all $t \in TERM_L$

are not even valid according to the above formalism. The reason is that besides individual constants whose denotations are determined solely by the context world, there are also complex singular terms whose denotations depend on the relevant index world. That is, there is a distinction between $Bel(a, \hat{x}P(x)(t))$ and $\hat{x}Bel(a, P(x))(t)$, if $t \in TERM_L$, and between $\Box(P(\iota\hat{x}A))$ and $\hat{x}\Box(P(x))(\iota\hat{y}A))$. These differences show that the abstraction principle, $A[^t/_x] \equiv \hat{x}A(t)$ is not valid.[12] Because universal instantiation is not valid in the above semantics, we can no longer derive the principle that any two co-referential singular terms can be substituted for each other without change in truth value, although the substitution principle of identicals, (SI), $\forall x \forall y[x = y \rightarrow (A(x) \leftrightarrow A(y))]$, is valid.

Although we don't give up the intuition that objects can only be identical to themselves, and to nothing else, it is easy to see that our semantics can account for *contingent identity*; for each object d and world w where d does not exist there might be two counterpart functions, c and c', such that $c_w(d) \neq c'_w(d)$. Note also that the definable counterpart relations don't need to be symmetric or transitive. It is possible for a counterpart function c in C that if $d \in D(w)$ and $c_{w'}(d) = d'$, that $c_w(d') \neq d$, and that if $c_{w'}(d) = d'$ and $c_{w''}(d') = d''$, it still doesn't have to hold that $c_{w''}(d) = d''$.

Of course, it is possible to make the extra assumption that each counterpart function gives rise to an equivalence relation. This can be done by assuming that each counterpart function c has to satisfy the following constraints: $\forall w, w', d, d'$: (i) $c_{w'}(c_w(d)) = c_w(d)$, and (ii) if $d \in D'(w)$ and $c_{w'}(d) = d'$, then $c_w(d') = d$. Both Lewis (1986), and Stalnaker (1986) have argued, however, that the (resulting) counterpart relation should not be transitive.

The main reason why I have used the counterpart theory here is that with its help we can account for certain problematic *de re* belief attributions. Ralph can believe of Ortcutt that he is a spy, and can believe of Ortcutt that he is not a spy, because in the different cases different representatives of Ortcutt in the belief worlds of Ralph were picked out; one

[12]But note that (EG) and (UI) are two aspects of the same principle; we can derive the one from the other by contraposition and double negation elimination. And, as noted by Thomason & Stalnaker (1968), also the principle of abstraction is closely related with the others in that the failure of (UI) and of the principle of abstraction are two sides of the same coin.

representative by counterpart function c, and another by counterpart function c'. Which counterpart function is relevant for communication depends on pragmatics, but I have assumed that semantically speaking we existentially quantify over counterpart functions.

Making use of free logic makes it possible that singular terms have no denotation in a particular world. There are two reasons why singular terms might fail to have a denotation in our semantics; First, because, for instance, the predicate P of the iota term $\iota\hat{x}P$ has an empty denotation in the world under consideration, or because a name has no causal origin and does not refer in that world. The second reason might be that an individual might have no counterpart in the world under consideration according to the relevant counterpart function. Once we allow for terms having no denotation in a world, we must decide how to interpret formulae containing such terms in that world. The easiest way to solve this problem is simply to assume that such atomic formulae are false. I haven't made this decision, however, because I want to account for an intuition proponents of situation semantics have argued possible worlds semantics cannot account for. They claim that in possible worlds semantics we cannot account for the fact that in some contexts we might truly and appropriately say *Mary believes that John walks*, although this is not the case for *Mary believes that John walks and Bill talks or doesn't talk*. The reason is that Mary might have no beliefs *about* Bill at all. Unfortunately, so they claim, the embedded sentences of the two belief attributions express the same proposition according to possible worlds semantics, so the difference between the two sentences cannot be accounted for in this framework. But of course, once the question is one of *aboutness*, we should check in our possible world semantics whether the analysis of *de re* belief attributions can account for this difference. And it can! If Mary has no belief about Bill, there will be no way in which Mary is *acquainted* with Bill, and the embedded sentence of the second clause will not be true in any of Mary's belief worlds with respect to any counterpart function. It follows that the belief attribution cannot be counted as being true.

Although our two-dimensional counterpart theory can account for most problems we have discussed in Chapter 1, at least some problems of *de re* belief attributions recently discussed by a number of authors cannot be handled appropriately. The discussion of these problems I will leave to another occasion, however.

Appendix B

Context Change Theory

Syntax

The syntax of the language **L** is the same as that of standard first-order predicate logic without individual constants. The lexicon of **L** has the following ingredients:

(i) basic symbols: $\neg, \wedge, \exists, \forall,), (, =$;

(ii) individual variables: $VAR_L = \{x_1, x_2, ...\}$;

(iii) for every $n \geq 0$, the set of n-place predicate constants: $PRED^n_L = \{P^n_1, P^n_2, ...\}$

The language **L** is defined by the terms of **L**, $TERM_L$, which is equal to VAR_L, and by the formulae of **L**, given by the following definition:

The set of formulae of **L**, $FORM_L$, is the smallest set such that:

(i) if $t_1...t_n \in TERM_L$ and $P \in PRED^n_L$, then $Pt_1...t_n \in FORM_L$;

(ii) if $t_1, t_2 \in TERM_L$, then $t_1 = t_2 \in FORM_L$;

(iii) if $A \in FORM_L$, then $\neg A \in FORM_L$;

(iv) if $A, B \in FORM_L$, then $A \wedge B \in FORM_L$;

(v) if $x \in VAR_L$, then $\exists x \in FORM_L$.

Disjunction and implication can be treated syncategorematically, by having '$(A \vee B)$' and '$(A \rightarrow B)$' stand for '$\neg(\neg A \wedge \neg B)$' and '$\neg(A \wedge \neg B)$', respectively. A formula like '$\exists x A$' is analysed as the conjunction of '$\exists x$' and 'A', and '$\forall x A$' as an abbreviation for '$\neg\exists x \neg A$'.

Semantics

Models are triples $\langle D, W, I\rangle$, where D is a non-empty set of objects, W a non-empty set of possible worlds, and I the intensional interpretation function that maps n-ary relations to a function from worlds to sets of n-tuples of objects. The set G of *partial assignments* associated with D and L is $\bigcup\{D^X |\ X \subseteq VAR_L\}$.

An *information state* S with domain X is a set of assignment-world pairs ($S \subseteq G \times W$) such that for all $\langle g, w\rangle$ that are elements of S, it holds that X is the domain of g. I will say that in these cases X is the domain of S, $D(S) = X$. I will use the following notational conventions with assignments g and h, objects d, variables x and y, and worlds w, where $x \notin dom(g)$ and for no $\langle g, w\rangle \in S:\ x \in dom(g)$:

- $g[x]h$ iff $dom(h) = dom(g) \cup \{x\}\ \&\ \forall y \in dom(h)[y \neq x \rightarrow h(y) = g(y)]$
- $S[x] \stackrel{def}{=} \{\langle h, w\rangle |\ \exists g:\ \langle g, w\rangle \in S\ \&\ g[x]h\}$
- $S[x := d] \stackrel{def}{=} \{\langle h, w\rangle | \exists g:\ \langle g, w\rangle \in S\ \&\ g[x]h\ \&\ h(x) = d\}$

The elements of $(G \times W)$ are ordered by $\leq$: $\langle g, w\rangle \leq \langle h, w'\rangle$ iff $w = w'$ and $g \subseteq h$. This ordering relation carries over to information states S and $S' : S \leq S'$ iff for every $\alpha \in S$: there is a $\beta \in S' :\ \alpha \leq \beta$.

For the interpretation rule of negation, I introduce $\alpha < S$, saying that α has an *extension* in S, which is the case iff $D(\{\alpha\}) \subseteq D(S)\ \&\ \exists \beta \in S : \alpha \leq \beta$. *Subtracting* state S' from state S, $S - S'$, will leave us with those elements of S that have no extension in $S' : S - S' = \{\alpha \in S |\ \neg(\alpha < S')\}$.

The notation '$G(S)$' will be used to give us the set of assignment functions in S:

- $G(S) = \{g \in G |\ \exists w \in W :\ \langle g, w\rangle \in S\}$

If $\langle g, v\rangle$ is an assignment-world pair, $w(\langle g, v\rangle) = v$.

Now I can give a recursive definition of the context change potential $[[A]]$ $[\subseteq \wp(G \times W) \times \wp(G \times W)]$ of formulae A of **L**:

$1a)$ $[[Px_1...x_n]](S) = \{\alpha \in S |\ \langle ||x_1||^\alpha, ..., ||x_n||^\alpha\rangle \in I_{w(\alpha)}(P)\}$, if $\forall x_i :\ 1 \leq i \leq n :$ $\forall \alpha \in S :\ ||x_i||^\alpha$ is defined, undefined otherwise

$(1b)$ $[[x_1 = x_2]](S) = \{\alpha \in S |\ ||x_1||^\alpha = ||x_2||^\alpha\}$, if $\forall x_i :\ 1 \leq i \leq 2 :$ $\forall \alpha \in S :\ ||x_i||^\alpha$ is defined, undefined otherwise

$(1c)$ $[[\exists x]](S) = \{\langle h, w\rangle |\ \exists g:\ \langle g, w\rangle \in S\ \&\ g[x]h\}\ \ (= S[x])$ if $\forall g \in G(S):\ x \notin dom(g)$, undefined otherwise

The (static) term evaluation used in (1a) and (1b) is defined by:

$$||x||^{g,w} = \ g(x), \text{ if } x \in dom(g), \text{ undefined otherwise}$$

Given the induction step, I assume that $[[A]](S)$ and $[[B]](S)$ have already been defined (for given formulae A and B and information states S) and give the following:

$$\begin{array}{lll} (2) \ [[\neg A]](S) & = & S - [[A]](S) \\ & = & \{\alpha \in S |\ \neg\exists\beta[\alpha \leq \beta \ \&\ \ \beta \in [[A]](S)]\} \\ (3) \ [[A \wedge B]](S) & = & [[B]]([[A]](S)) \end{array}$$

Now I can define the most important semantic concepts. A formula A is *acceptable in* S, $S \models_d A$, iff S is a substate of $[[A]](S)$, in the sense that every $\alpha \in S$ can be extended to a $\beta \in [[A]](S)$ such that $\alpha \leq \beta$. A is *accepted in* S, $S \models_s A$, iff $S = [[A]](S)$. A *entails* B, $A \models_{d/s} B$, iff for all $S:\ [[A]](S) \models_{d/s} B$.

Appendix C

Pronouns as referential expressions

To give some more content to the suggestions made in section 2.3, I will now define a syntax and semantics for a formal language, which will ultimately serve to provide a semantics for natural language expressions. However, I will not give systematic translation rules from natural language to this formal language, but rely, instead, on the reader's willingness to translate/represent natural language expressions in obvious ways.

Syntax

The language **L** has the following symbols:[1]

(i) basic symbols: $\neg$, $\wedge$, *Det*, *ADV*,), (, $\hat{}$, η, ι;
(ii) a denumerable set of individual variables: $VAR_L = \{x, y, ...\}$;
(ii) a denumerable set of discourse markers: $DR_L = \{r_1, r_2, ...\}$;
(iii) a denumerable set of indices: $Ind_L = \{n, m, ...\}$;
(iv) for every $n \geq 0$, a denumerable set of primitive n-place predicates.

The language **L** is defined in accordance with the following definition of the terms and complex expressions of **L**:

The set of *terms* of **L**, $TERM_L$, is equal to $VAR_L \cup DR_L \cup CSTERMS_L$, where $CSTERMS_L$ is the set of complex singular terms of **L** to be defined below.

The complex expressions of **L** are sentences, complex predicates, or complex singular terms. These sets are defined *simultaneously* as follows:

(a) Sentences:

(i) If $t_1...t_n$ are terms and P is an n-place predicate, then $Pt_1...t_n$ is a sentence.

[1] Where *Det* is any determiner and *ADV* any kind of adverb of quantification.

(ii) If A is a sentence, then $\neg A$ is a sentence.
(iv) If A and B are sentences, then $A \wedge B$ and $ADV(A, B)$ are sentences.
(v) If P is a one-place predicate, then $\exists P$ is a sentence.
(vi) If A and B are sentences, and x a variable, then $Det_x(A, B)$ is a sentence.

(b) Complex predicates:

(i) If A is a sentence and x is a variable, then $\hat{x}A$ is a one-place predicate.

(c) Complex singular terms:

(i) If P is a one-place predicate, and r is a discourse marker,
then $\iota r P$ is a complex singular term,
(ii) If P is a one-place predicate, r is a discourse marker, and n is an index,
then $\eta r_n P$ is a complex singular term.

There are no other complex expressions.

The formulae '$A \vee B$', '$A \rightarrow B$', and '$\forall A$' will be abbreviations for '$\neg(\neg A \wedge \neg B)$', '$\neg(A \wedge \neg B)$', and '$\neg\exists\neg A$', respectively.

Thus, the syntax of our language **L** is just like (a version of) ordinary predicate logic, the only differences being that (i) if A and B are sentences, then $Q(A, B)$ is a sentence too, where Q is either a determiner with a variable, or an adverb of quantification; (ii) if A is a sentence, then $\hat{x}A$ is a one-place predicate; and (iii) we allow for the existence of complex singular expressions in the form of *iota* and (indexed) *eta* terms. Iota terms are used to represent definite descriptions, and (indexed) eta terms are used to represent specifically-used indefinites to which we can refer back by singular pronouns that are not used descriptively. I will assume that in a discourse, each occurrence of a (specifically-used) indefinite should be represented by an eta term with a different index. The abstraction operator is added to our language to account for the *scope* of (complex singular) terms (see Thomason & Stalnaker, 1968), and to analyse anaphoric pronouns c-commanded by coreferential singular terms.

Semantics

The semantics is given relative to intensional models of the following form: $\langle W, D, *, C, I\rangle$, where W is a set of possible worlds; D a set of individuals figuring as our domain; $*$ a special object that is not an element of D; C a denumerable set of reference contexts, total functions from indices to $D \cup \{*\}$ such that for each $\vec{m} \in Ind^n$ and $\vec{d} \in D^n$ there is a $c \in C$ such that $\langle c(m_1), ..., c(m_n)\rangle = \langle d_1, ..., d_n\rangle$;[2] and I the interpretation function that assigns

[2]This constraint is needed to account for quantification and donkey sentences.

to n-ary predicates a function from worlds to a relation between n individuals. I assume that for each w in W, there is a distinguished c in C in the sense explained above. For the semantics we also need the set of assignment functions G that assign individuals of the domain of the model to variables and discourse referents.

Before we can give the actual truth definition, we first have to provide some definitions. The first is for the notion '$Upd(A, \langle w, c, g\rangle)$', which gives us the *partial* assignment function g enriched by the objects introduced by the terms used in A in possibility $\langle w, c, g\rangle$ under their respective variables.[3] The second is for the notion of *rigid* truth, '$[[A]]^{w,c,g} = 1$', which should not be confused with the actual, non rigid, truth definition. The last is to give the interpretation rules for complex predicates and terms. In the end, these notions have to be defined simultaneously, but as long as we ignore descriptive pronouns and presuppositions we can define $Upd(A, \langle w, c, g\rangle)$ as follows:[4,5]

- $Upd(\eta r_n P, \langle w, c, g\rangle) = \quad Upd(P, \langle w, c, g[^r/_{c(n)}]\rangle)$;
- $Upd(t, \langle w, c, g\rangle) = \quad g$, if t is a variable or discourse referent;
- $Upd(R(t_1, .., t_n), \langle w, c, g\rangle) = Upd(R, \langle w, c, Upd(t_n, .., Upd(t_1, \langle w, c, g\rangle)..)\rangle)$;
- $Upd(R, \langle w, c, g\rangle) = g$, if R is a primitive relation
  $= Upd(A, \langle w, c, g\rangle)$, if R is of the form $\hat{x}A$;
- $Upd(A \wedge B, \langle w, c, g\rangle) = \quad Upd(B, \langle w, c, Upd(A, \langle w, c, g\rangle)\rangle)$;
- $Upd(\exists P, \langle w, c, g\rangle) = Upd(P, \langle w, c, g\rangle)$;
- $Upd(\neg A, \langle w, c, g\rangle) = g$;
- $Upd(ADV(A, B), \langle w, c, g\rangle) = g$;
- $Upd(Det_x(A, B), \langle w, c, g\rangle) = g$.

Accordingly, each occurrence of an indefinite introduces to each possibility a unique and specific individual, intuitively its speaker's referent and formally the object that the reference context of the possibility assigns to the index of the eta term. If an indefinite is embedded under a negation, quantifier, or adverb of quantification, it doesn't introduce an

[3] In the case of an iota term, the unique individual, if any, that satisfies the descriptive material of the definite description; and in the case of an eta term, the speaker's referent of this occurrence of the term.

[4] Note that whereas in standard dynamic semantics the introductions of individuals/discourse referents and the evaluation of truth in a world are accounted for by *one* update function, I account for the two processes that are conceptually distinct by two separate definitions. Not only my definition of non-rigid truth of a sentence, that will be defined later, but also the phenomenon of pronominal contradiction shows, I believe, that this separate treatment is needed.

[5] I will give interpretation rules only for the sentences/constructions in the original fragment of DRT/FCS/DPL.

individual to the possibility with respect to which the embedded sentence is interpreted, just as in ordinary dynamic semantics.

We can define the notion of 'rigid truth' as follows (where $I_{w,c,g}(R) = I_w(R)$ if R is a primitive predicate, and $[Q]$ denotes the interpretation of Q):

- $[[R(t_1, ..., t_n)]]^{w,c,g} = 1$ iff $\langle [[t_1]]^{w,c,g}, ..., [[t_n]]^{w,c,g} \rangle \in I_{w,c,h}(R)$ where $h = Upd(t_n\langle w, c, Upd(t_{n-1}, ..., Upd(t_1, \langle w, c, g\rangle)...)\rangle)$;
- $[[A \wedge B]]^{w,c,g} = 1$ iff $[[A]]^{w,c,g} = 1$ and $[[B]]^{w,c,h} = 1$, where $h = Upd(A, \langle w, c, g\rangle)$;
- $[[\neg A]]^{w,c,g} = 1$ iff $\neg\exists c' \in C:\ [[A]]^{w,c',g} = 1$;
- $[[\exists P]]^{w,c,g} = 1$ iff $I_{w,c,g}(P) \neq \emptyset$;
- $[[ADV(A, B)]]^{w,c,g} = 1$ iff $[ADV](\{Upd(A, \langle w, c', g\rangle):\ c' \in C\ \&\ [[A]]^{w,c',g} = 1\}, \{Upd(A, \langle w, c', g\rangle):\ c' \in C\ \&\ [[A \wedge B]]^{w,c',g} = 1\})$;
- $[[Det_x(A, B)]]^{w,c,g} = 1$ iff $[Det\ ](\{d \in D:\ \exists c' \in C\ \&\ [[A]]^{w,c',g[x/d]} = 1\}, \{d \in D : \exists c' \in C\ \&\ [[A \wedge B]]^{w,c',g[x/d]} = 1\})$.

Next we can give the interpretation rules for complex predicates and terms:

- $I_{w,c,g}(\hat{x}A) = \{d \in D:\ [[A]]^{w,c,g[x/d]} = 1\}$
- $[[t]]^{w,c,g} = g(t)$, if t is a variable or discourse referent,
  $= d$, if $t = \iota r P$ and $I_{w,c,g}(P) = \{d\}$,
  $= d$, if $t = \eta r_n P$, $c(n) = d$ and $d \in I_{w,c,g}(P)$,
  $= *$ otherwise.

Finally we can define the notion of *truth of sentence* A in $\langle w, c, g\rangle$, $\langle w, c, g\rangle \models A$, in terms of the above notions, as follows:

- $\langle w, c, g\rangle \models A$ iff there is a $c' \in C$ such that $[[A]]^{w,c',g} = 1$.

Note that I assume that for the interpretation of the rigid truth of atomic clauses, the terms are interpreted independently of each other; the dynamic effect can be relevant only for the interpretation of the predicate. This seems to raise problems for sentences containing anaphoric pronouns that are c-commanded by coreferential singular terms, which include sentences with reflexive pronouns or those like *Mary loves her uncle*. Fortunately, we can solve these problems by assuming that the anaphoric relations in such sentences should be represented by means of the abstraction operator. *Mary loves her uncle*, for instance, can be represented by the following formula: $\hat{x}Love(x, \iota r\hat{y}Uncle - of(y, x))(m)$, on the assumption that constants are added to our language. A sentence containing an indefinite in a relative clause such as *A farmer who owns a donkey is beating it* seems to be more

problematic. But this sentence can also be interpreted if we represent the sentence by $\hat{x}Beat(x,s)(\eta r_n\hat{x}(Fx \wedge Own(x,\eta s_m D)))$, because the internal dynamic effect is assumed to be relevant to the interpretation of the predicate.

It is useful to make an explicit calculation of the truth-conditions of this sentence. The formula is (rigidly) true in $\langle w,c,g\rangle$ if the referent of $[[\eta r_n\hat{x}(Fx\wedge Own(x,\eta s_m D))]]^{w,c,g}$ is an element of $I_{w,c,h}(\hat{x}Beat(x,s))$, where $h = Upd(\eta r_n\hat{x}(Fx\wedge Own(x,\eta s_m D)),\langle w,c,g\rangle)$. By inspecting the definition of '$Upd(A,\langle w,c,g\rangle)$', one can see that $h = g[^r/_{c(n)},^s/_{c(m)}]$. If we assume that $c(n)$ is a farmer who owns a donkey, $c(n) \in I_{w,c,g}(\hat{x}(Fx\wedge Own(x,\eta s_m D)))$, and that $c(m)$ is a donkey, $c(m) \in I_w(D)$, the sentence is true in $\langle w,c,g\rangle$ iff $c(n) \in \{d \in D : [[Beat(x,s)]]^{w,c,g[^r/_{c(n)},^s/_{c(m)}][^x/_d]} = 1\}$. But this holds exactly if $\langle c(n),c(m)\rangle \in I_w(Beat)$, just as we want.

It is useful to determine explicitly when a formula containing a specifically-used indefinite and a conjunction is rigidly true. Consider, for example, the formula $Q(\eta s_n P) \wedge Rs$, interpreted in possibility $\langle w,c,g\rangle$. This conjunction is (rigidly) true, $[[Q(\eta s_n P) \wedge Rs]]^{w,c,g} = 1$, iff the first conjunct is rigidly true with respect to $\langle w,c,g\rangle$; and the second conjunct (rigidly) true with respect to $\langle w,c,g\rangle$, updated with the referents of the terms used in the first conjunct, i.e. $\langle w,c,Upd(Q(\eta s_n P),\langle w,c,g\rangle)\rangle$. First we determine whether the first conjunct is rigidly true, $[[Q(\eta s_n P)]]^{w,c,g} = 1$. According to the interpretation rule of atomic formulae this is the case when $[[\eta s_n P]]^{w,c,g} \in I_{w,c,g}(Q)$. By inspecting the interpretation rule for terms, this is the case iff the speaker's referent of $\eta s_n P$ in $\langle w,c,g\rangle$ is a P, and has the property denoted by Q. This in turn holds iff $c(n) \in I_{w,c,g}(P)$ and $c(n) \in I_{w,c,g}(Q)$, which is the case iff $c(n) \in (I_w(P)\cap I_w(Q))$, if P and Q are assumed to be primitive predicates. To be able to determine the truth value of the second conjunct, we first have to determine the assignment with respect to which the second conjunct has to be interpreted, i.e. $Upd(Q(\eta s_n P),\langle w,c,g\rangle)$. By inspecting the definition of $Upd(A,\langle w,c,g\rangle)$ we see that this is the same assignment as $Upd(\eta s_n P,\langle w,c,g\rangle)$; again on the assumption that P is a primitive predicate, this in turn is equivalent to $g[^s/_{c(n)}]$. The second conjunct is now (rigidly) true with respect to this enriched possibility, $[[Rs]]^{w,c,g[^s/_{c(n)}]} = 1$, iff $g[^s/_{c(n)}](s) \in I_{w,c,g[^s/_{c(n)}]}(R)$. This latter condition holds iff $c(n) \in I_w(R)$, if R is assumed to be a primitive predicate. As a result, the whole conjunction is true in possibility $\langle w,c,g\rangle$ iff $c(n) \in (I_w(P) \cap I_w(Q) \cap I_w(R))$, which intuitively means that the speaker's referent of the occurrence of the indefinite has to have all three properties denoted by P, Q, and R.

As explained in section 2.3, a treatment of negation and (adverbial) quantifiers as 'intensional' operators, i.e. shifters of reference contexts, allows us to account for the universal effect of donkey sentences. One way to spell out such an account is to have the two arguments fronted by an implicit adverb of quantification, $Always(Own(\eta r_n\hat{x}Fx,\eta s_m\hat{y}Dy),Beat(r,s))$. Notice that this analysis accounts only for the *unselective* reading of a donkey sentence; i.e. in making use of quantification over farmer-donkey pairs only. This leads to the question whether we can also account for the *selective*, or *asymmetric*, reading of donkey sentences, where we seem to quantify, for instance, only over farmers. Accounting for the asymmetric

readings turns out, however, to be simple, if we quantify, as usual, over *equivalence classes* of cases (cf. Chierchia (1992) and Dekker (1993)) – in our case, world/reference context pairs. We can account for this formally by indexing the adverb/determiner with some but not all variables introduced by the restrictor/antecedent and interpreting such (adverbial) quantificational sentences as follows:

$$[[Qr_1, ..., r_n(A, B)]]^{w,c,g} = 1 \quad \text{iff}$$
$$[Q](\{\langle h(r_1), ..., h(r_n)\rangle : \exists c' : [[A]]^{w,c',g} = 1 \ \&\ Upd(A, \langle w, c', g\rangle) = h\},$$
$$\{\langle h(r_1), ..., h(r_n)\rangle : \exists c' : [[A \wedge B]]^{w,c',g} = 1 \ \&\ Upd(A \wedge B, \langle w, c', g\rangle) = h\}).$$

Our notion of 'rigid truth' corresponds roughly to the notion of truth assumed by Chastain (1975), Donnellan (1978), and Fodor & Sag (1982): if the indefinite *an S* in *An S is P* is used specifically, the sentence is predicted to be false if the specific speaker's referent of the indefinite doesn't have property *P*, even though another individual with property *S* does. However, to account for the intuition that such sentences have only *existential* truth conditions, we have defined our non-rigid notion of truth in such a way that we always abstract away from speaker's reference when analysing clauses containing indefinites. We have said that sentence A is true in $\langle w, c, g\rangle$, $\langle w, c, g\rangle \models A$ iff there is a $c' \in C$ such that $[[A]]^{w,c',g} = 1$. Thus, for the truth of each individual sentence we existentially quantify over reference contexts, thereby making the specificity of the indefinites truth-conditionally irrelevant for the sentence in which the indefinite occurs. On the other hand, the individuals introduced by this sentence, the referents of the pronouns of later sentences, depend exclusively on the actual reference context. By means of this truth definition we can thus make a distinction between *relative* and *personal* pronouns.

Appendix D

The Triviality result

According to any standard analysis of probability, the result of successive conditionalisation on two statements is the same as that of conditionalising once on the conjunction of those statements. This can be illustrated by the following example:

Suppose an unbiased coin is tossed two times. The value our subjective probability function P will assign to heads of the second toss, $P(h_2)$, will be $1/2$. After we learn that at least one of the two tosses yielded heads, our probability assigned to h_2 will be $P(h_2/h_1 \vee h_2) = (1/2)/(3/4) = 2/3$. Let's call the new resulting probability function P'. If we learn that the two tosses did not both yield heads, we conditionalise P' by $\neg(h_1 \wedge h_2), P'(h_2/\neg(h_1 \wedge h_2)) = P'(h_2 \wedge \neg(h_1 \wedge h_2))/P'(\neg(h_1 \wedge h_2)) = (1/3)/(2/3) = 1/2$. So the probability of h_2 according to the probability function P'' resulting after two times conditionalising is $1/2$. The same results if we conditionalise once on the conjunction of those two statements, $P''(h_2) = P(h_2/((h_1 \vee h_2) \wedge \neg(h_1 \wedge h_2)))$.

Suppose now that $(B/A)/C$ makes sense as a statement to which a probability function P can be applied that obeys the usual conditions. That is, let us make the crucial assumption of Stalnaker (1970a), i.e. that '/' obeys conditionalisation *and* is a connective with a context independent meaning. Then it follows that for any probability function $P, P((B/A)/C) = P(B/A \wedge C)$, if $P(A \wedge C) \neq 0$. This then, really is (CSH). The condition (CSH) can be proven on the following definition of a subfunction, and the four assumptions below:

Definition: A *subfunction*, P_A, is a function defined for any probability function P and proposition A such that $P(A) \neq 0$ as follows: $P_A(B) \stackrel{def}{=} P(B/A)$

Assumptions:

(0) $P(B/A)$ is defined only when $P(A) \neq 0$,

(1) If $P(A) \neq 0, P(A > B) = P(B/A)$, Stalnaker's hypothesis,

(2) Any subfunction is a probability function,

(3) The conditional has a fixed interpretation, $A > B$ expresses the same proposition in all contexts/probability functions.

The assumption that the conditional has a fixed interpretation is used by Lewis in the following form: ">" means the same in P and in P_C, for any C and P. On the basis of these assumptions, we can prove (CSH):

$$(CSH) \quad P(A > B/C) = P(B/A \wedge C), \text{ if } P(A \wedge C) \neq 0.$$

Lewis (1975) derived the triviality result from (CSH), which follows from the assumptions made in Stalnaker (1970a). Stalnaker (1976a) showed how (CSH) follows from his assumptions:

$$\begin{array}{lll}
P_C(A \wedge B) = P_C(A) \times P_C(B/A) & & \text{(by axioms of } P) \\
= P_C(A) \times P_C(A > B), \text{ if } P_C(A) \neq 0, & & \text{(by (1))} \\
\text{(a)} \quad = P(A/C) \times P(A > B/C), \text{if } P(A \wedge C) \neq 0, & & \text{(by subfunction and (3))} \\
P_C(A \wedge B) = P(A \wedge B/C) & & \text{(by subfunction)} \\
\text{(b)} \quad = P(A/C) \times P(B/A \wedge C) & & \text{(by axioms of } P) \\
\text{If } P(A \wedge C) \neq 0, P(A > B/C) = P(B/A \wedge C) & & \text{(by (a) and (b))}
\end{array}$$

For Lewis' triviality proof, we first prove an independence property, saying that $P(A \wedge B) = P(A) \times P(B)$. The proof of this independence property is based on (CSH) and the following two standard assumptions:

(4) $P(A) = P(A/B) \times P(B) + P(A/\neg B) \times P(\neg B)$, if $0 \neq P(B) \neq 1$,
expansion by cases

(5) if $P(B) \neq 0$, then (a) if $A \models B$, then $P(B/A) = 1$, and
(b) if $A \models \neg B$, then $P(B/A) = 0$.

The essential step to prove that $P(C \wedge A) = P(C) \times P(A)$, is to show that on assuming (CSH) one can derive that $P(C/A) = P(C)$:

$$\begin{array}{lll}
P(C/A) = \text{(4)} & P(A > C/C) \times P(C) + P(A > C/\neg C) \times P(\neg C), & \\
& \text{if } P(C) \neq 0 \neq P(\neg C) & \\
= \text{(CSH)} & P(C/C \wedge A) \times P(C) + P(C/\neg C \wedge A) \times P(\neg C), & \\
& \text{if } P(C \wedge A) \neq 0 \neq P(\neg C \wedge A) & \\
= \text{(5)} & 1 \times P(C) + 0 \times P(\neg C) & \\
= & P(C) &
\end{array}$$

By conditionalisation: $P(C \wedge A) = P(C/A) \times P(A) = P(C) \times P(A)$.

So, from Stalnaker's hypothesis together with the assumption that the conditional has a fixed interpretation, it follows that $P(A \wedge (A > C)) = P(A) \times P(A > C)$. It is thus predicted that $P(A)$ and $P(A > C)$ are probabilistically independent of each other. Assuming $P(A \wedge B) = P(A) \times P(B)$ for any A and B, we can prove **Lewis' triviality result:**

If $P(A) \neq 0, P(A > B) = P(B/A)$, any probability function P that uses conditionalisation can assign to at most two pairwise incompatible propositions a non-zero probability.

Proof: Let C, D and E be three pairwise incompatible propositions with non-zero probability. Assume $A = C \vee D$ and Stalnaker's hypothesis. By the incompatibility of C and D it follows that $P(A \wedge C) = P(C)$, and by Stalnaker's hypothesis it follows that $P(A \wedge C) = P(A) \times P(C)$, because it is predicted that A and C are probabilistically independent. As a result it is predicted that $P(A) \times P(C) = P(C)$. Thus, $P(A) = 1$ and $P(\neg A) = 0$. But this is impossible because $P(\neg A) \geq P(E)$, which has by hypothesis a non-zero probability.

So even without assuming that $A > C$ obeys Stalnaker's logic $C2$, Stalnaker's hypothesis, $P(A > B) = P(B/A)$, cannot be made.

Bibliography

[1] Abusch, D. (1993), "The scope of indefinites", *Natural Language Semantics*, 2, pp. 83-135.

[2] Adams, E.W. (1965), "The logic of conditionals", *Inquiry*, 8, pp.166-197.

[3] Adams, E.W. (1970), "Subjunctive and indicative conditionals", *Foundations of Language*, 6, pp. 89-94.

[4] Adams, E.W. (1976), "Prior probabilities and counterfactual conditionals", In: W.L. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol I*, Reidel, Dordrecht, pp. 1-22.

[5] Aloni, M. (2001), *Quantification under Conceptual Covers*, Ph.D. dissertation, University of Amsterdam.

[6] Asher, N. (1987), "A typology for attitude verbs and their anaphoric properties", *Linguistics and Philosophy*, 10, pp. 125-197.

[7] Bäuerle R. and M. J. Cresswell, (1984), "Propositional attitudes", In: D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*, Vol. IV, Reidel, Dordrecht, pp. 491-512.

[8] Beaver, D. (1995), *Presupposition and Assertion in Dynamic Semantics*, Ph.D. thesis, CCS, Edinburgh.

[9] Beaver, D. (2001), *Presupposition and Assertion in Dynamic Semantic*, Studies in Logic, Language and Information, CLSI Publications, Stanford.

[10] Belnap, N.D. (1970), "Conditional assertion and restricted quantification", *Nous*, 4, pp. 1- 12.

[11] Berg, M.H. van den (1996), *The Internal Structure of Discourse*, Ph.D. dissertation, University of Amsterdam.

[12] Bochvar, D.A. (1939), "Ob odsom trehznachom iscislenii i ego primeneii k analizu paradoksov klassicskogo rassirennoga funkcional 'nogo iscislenija' ", *Mathematiciskij*

sbornik, 4. (translated by M. Bergmann as "On a three-valued calculus and its application to the analysis of the paradoxes of the classical extended functional calculus", *History and Philosophy of Logic*, 2 (1981), pp. 87-112.)

[13] Bratman, M.E. (1987), *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge.

[14] Burge, T. (1979), "Individualism and the mental", In: P. French et al. (eds.), *Midwest Studies in Philosophy, 4, Studies in Epistemology*, University of Minnesota Press, Minneapolis, pp. 73-122.

[15] Buridan, J. (1350), *Sophismata*, trans. T.K. Scott as *Sophisms on Meaning and Truth*, Appleton-Century-Crofts, New York, 1966.

[16] Chastain, C. (1975), "Reference and context", In: K. Gunderson (ed.), *Minnesota Studies in the Philosophy of Science, vol. VII- Language, Mind, and Knowledge*, University of Minnesota Press, Minneapolis, pp. 194-269.

[17] Chierchia, G. (1992), "Anaphora and Dynamic Binding", *Linguistics and Philosophy*, 12, pp. 111-183.

[18] Church, A. (1954), "Intensional isomorphism and identity of belief", *Philosophical Studies*, 5, pp. 65-73.

[19] Cohen, P.R. & H.J. Levesque (1990), "Intention is choice with commitment", *Artificial Intelligence*, 42, pp. 213-261.

[20] Cooper, R.H. (1979), "The interpretation of pronouns", In: F. Heny and H.S. Schnelle (eds.), *Selections from the third Groningen round table, Syntax and Semantics*, 10, Academic Press, New York, pp. 61-92.

[21] Cresswell, M. (1973), *Logics and Languages*, Methuen, London.

[22] Cresswell, M. and A. von Stechow, (1982), "De re belief generalised", *Linguistics and Philosophy*, 5, pp. 503-535.

[23] Crimmins, M and J. Perry, (1989), "The prince and the phone booth: reporting puzzling beliefs", *Journal of Philosophy*, 86, pp. 685-711.

[24] Deemter, K. van (1991), *On the Composition of Meaning*, Ph.D. dissertation, University of Amsterdam.

[25] Dekker, P. (1993), *Transsentential Meditations, Ups and downs in dynamic semantics*, Ph.D. dissertation, University of Amsterdam.

[26] Dekker, P. (1994), "Predicate logic with anaphora", In: R. Cooper and J. Groenendijk, *Integrating Semantic Theories II*, Dyana-2, Deliverable R2.1.B.

[27] Dekker, P. (1997), "On first order information exchange", In: A. Benz & G. Jäger, *Proceedings of Mundial'97. Munich Workshop on the Formal Semantics and Pragmatics of Dialogue*, CIS, München, pp. 21-39.

[28] Dekker, P. and R. van Rooy (1998), "Hob-Nob sentences, and Hob-Nob situations", In: R. Cooper & T. Gamkrelidze (eds.), *Proceedings of the 2nd Tbilisi symposium on Language, Logic and Information*, Tbilisi, pp. 86-97.

[29] Dennett, D.C. (1969), *Content and Consciousness*, Routledge and Kegan Paul, London.

[30] Does, J. van der. (1994), "Formalising E-type logic", In: P. Dekker and M. Stokhof (eds.), *Proceedings of the Ninth Amsterdam Colloquium*, Amsterdam, pp. 229-248.

[31] Donnellan, K. (1966), "Reference and definite descriptions", *Philosophical Review*, 75, pp. 281-304.

[32] Donnellan, K. (1970), "Proper names and identifying descriptions", *Synthese*, 21, pp. 3-31.

[33] Donnellan, K. (1974), "Speaking of nothing", *Philosophical Review*, 83, pp. 3-32.

[34] Donnellan, K. (1978), "Speaker reference, descriptions, and anaphora", In: P. Cole (ed.), *Syntax and Semantics, vol. 9: Pragmatics*, Academic Press, New York, pp. 47-68.

[35] Dretske, F.L. (1970), "Epistemic operators", *The Journal of Philosophy*, 67, pp. 1007- 1023.

[36] Dretske, F.L. (1981), *Knowledge and the Flow of Information*, MIT Press, Cambridge, Massachusetts.

[37] Edelberg, W. (1986), "A new puzzle about intentional identity", *Journal of Philosophical Logic*, 15, pp. 1-25.

[38] Edelberg, W. (1992), "Intentional identity and the attitudes", *Linguistics and Philosophy*, 15, pp. 561-596.

[39] Edelberg, W. (1995), "A perspectival semantics for the attitudes", *Nous*, 29, pp. 316-342.

[40] Eijck, J. van and G. Ceparello, (1994), "Dynamic modal predicate logic", In: M. Kanazawa and C. Pinon (eds.), *Dynamics, Polarity and Quantification*, CSLI, Stanford, pp. 251-276.

[41] Evans, G. (1973), "The causal theory of names", *Proceedings of the Aristotelian Society*, Supplementary Volume 47, pp. 187-208.

[42] Evans, G. (1977), "Pronouns, quantifiers and relative clauses (1)", *The Canadian Journal of Philosophy*, 7, pp. 467-536.

[43] Evans, G. (1979), "Reference and contingency", *The Monist*, 62, pp. 161-189.

[44] Evans, G. (1981), "Understanding demonstratives", In: H. Parret and J. Bouveresse (eds.), *Meaning and Understanding*, De Gruyter, Berlin, pp. 280-303.

[45] Evans, G. (1982), *Varieties of Reference*, Oxford University Press, Oxford.

[46] Fagin, R. and J. Halpern, (1988), "Belief, awareness and limited reasoning", *Artificial Intelligence*, 34, pp. 39-76.

[47] Fernando, T. (1994), "Generalised quantifiers as second-order programs - 'dynamically' speaking, naturally", In: P. Dekker and M. Stokhof (eds.), *Proceedings of the Ninth Amsterdam Colloquium*, ILLC, Amsterdam, pp. 287-3000.

[48] Fernando, T. (1997), "Are context change potentials functions?", In: H. Kamp and B. Partee (eds.), *Context in the Analysis of Linguistic Meaning*, Stuttgart/Prague, pp. 129-152.

[49] Fine, K. (1975), "Review of Lewis's 'Counterfactuals' ", *Mind*, 84, pp. 451-458.

[50] Fintel, K. von (1994), *Restrictions on Quantifier Domains*, Ph.D. dissertation, University of Massachusetts, Amherst.

[51] Fodor, J.A. (1987), *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, MIT Press, Cambridge, Mass..

[52] Fodor, J.D. and I.A. Sag, (1982), "Referential and quantificational indefinites", *Linguistics and Philosophy*, 5, pp. 355-398.

[53] Fraassen, B.C. van (1973), "Values and heart's command", *Journal of Philosophy*, 70, 5-19.

[54] Fraassen, B.C. van (1974), "Hidden variables and conditional logic", *Theoria*, 40, pp. 176-190.

[55] Fraassen, B.C. van (1976), "Probabilities of conditionals", In: W.L. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Volume I*, Reidel, Dordrecht, pp. 261-308.

[56] Fraassen, B.C. van, (1977), "The only necessity is verbal necessity", *Journal of Philosophy*, 74, pp. 71-85.

[57] Fraassen, B.C. van, (1979), "Propositional attitudes in weak pragmatics", *Studia Logica*, 76, pp. 365-374.

[58] Frank, A. (1997), *Context Dependence in Modal Constructions*, Ph.D. dissertation, University of Stuttgart.

[59] Frege, G. (1892), "Über Sinn und Bedeutung", *Zeitschrift für Philosophie und philosophische Kritik*, 50, pp. 25-50.

[60] Gärdenfors, P. (1982), "Imaging and conditionalisation", *Journal of Philosophy*, 79, pp. 747-760.

[61] Gärdenfors, P. (1988), *Knowledge in Flux, Modeling the Dynamics of Epistemic States*, MIT Press, Cambridge Mass..

[62] Gärdenfors, P. and D. Makinson, (1994), "Nonmonotonic inference based on expectations", *Artificial Intelligence*, 65, pp. 197-245.

[63] Gazdar, G. (1979), *Pragmatics: Implicatures, Presuppositions, and Logical Form*, Academic Press, New York.

[64] Geach, P. (1962), *Reference and Generality, Cornell University Press*, Ithaca.

[65] Geach, P. (1967), "Intentional identity", *Journal of Philosophy*, 64, pp. 627-632.

[66] Geenhoven, V. van (1996), *Semantic Incorporation and Indefinite Descriptions*, Ph.D. Dissertation, University of Tübingen. (SfS-Report-03-96).

[67] Gerbrandy, J. (1999), *Bisimulations on Planet Kripke*, Ph.D dissertation, University of Amsterdam.

[68] Gettier, E. (1963), "Is justified true belief knowledge?", *Analysis*, 6, pp. 121-123.

[69] Geurts, B. (1995), *Presupposing*, Ph.D. Dissertation, University of Stuttgart.

[70] Geurts, B. (1998), "Presuppositions and anaphors in attitude contexts", *Linguistics and Philosophy*, 21, pp. 545-601.

[71] Gibbard, A. (1980), "Two recent theories of conditionals", In: W. L. Harper et al. (eds), *Ifs*, Reidel, Dordrecht, pp. 211-247.

[72] Gibbard, A. and W.L. Harper (1978), "Counterfactuals and two kinds of expected utility", In: C. Hooker et al. (eds.), *Foundations and Applications of Decision Theory*, Western Ontario Series in the Philosophy of Science, Vol. 1, Reidel, Dordrecht, pp.125-162.

[73] Grice, H.P. (1989), *Studies in the Way of Words*, Harvard University Press, Cambridge.

[74] Groenendijk, J. and Stokhof, M. (1982), "Semantic analysis of Wh-complements", *Linguistics and Philosophy*, 5, pp. 175-233.

[75] Groenendijk, J. and M. Stokhof (1990), "Dynamic Montague Grammar", In: L. Kalman and L. Polos (eds.), *Papers from the second symposium on logic and language*, Akademia Kiado, Budapest, pp. 3-48.

[76] Groenendijk, J. and M. Stokhof (1991), "Dynamic predicate logic", *Linguistics and Philosophy*, 14, pp. 39-100.

[77] Groenendijk, J, et al. (1996), "Coreference and modality", In: S. Lappin (ed.), *Handbook of Contemporary Semantic Theory*, Blackwell, Oxford, pp. 179-214.

[78] Groenendijk, J. et al. (1997), "Coreference and modality in the context of multi-speaker discourse", In: H. Kamp and B. Partee (eds.), *Context in the Analysis of Linguistic Meaning*, Stuttgart/Prague.

[79] Grove, A.J. (1988), 'Two Modellings for Theory Change', *Journal of Philosophical Logic*, 17, 157-170.

[80] Haas Spohn, U. (1986), "Zur interpretation der Einstellungszuschreibungen", *Sonderforschungsbereich 99*, University of Konstanz.

[81] Haas-Spohn, U. (1994), *Versteckte Indexikalitat und subjective Bedeutung*, Ph.D. Dissertation, University of Tübingen.

[82] Hajek, A. and N. Hall, (1994), "The hypothesis of the conditional construal of conditional probabilities", In: E. Eells and B. Skyrms (eds), *Probability and Conditionals*, Cambridge University Press, Cambridge, pp. 75-111.

[83] Hansson, S. O. (1989), "A new semantical approach to the logic of preference", *Erkenntnis*, 31, 1-42.

[84] Harper, W.L. (1976a), "Rational belief change, Popper functions, and counterfactuals", In: W. L. Harper & C. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. I*, Reidel, Dordrecht, pp. 73-112.

[85] Harper, W.L. (1976b), "Ramsey test conditionals and iterated belief change (A response to Stalnaker)", In: W.L. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Vol I, Reidel, Dordrecht, pp. 117-136.

[86] Harper, W.L. (1977), "Rational conceptual change", In: *PSA 1976. East Lansing Mich., Philosophy of Science Association*, 2, pp. 462-135.

[87] Harper, W.L. (1980), "A sketch of some recent developments in the theory of conditionals", In: W.L. Harper et al. (eds.), *Ifs*, Reidel, Dordrecht, pp. 3-38.

[88] Hazen, A. (1979), "Counterpart theoretic semantics for modal logic", *The Journal of Philosophy*, 76, pp. 319-338.

[89] Heim, I. (1982), *The Semantics of Definite and Indefinite Noun Phrases*, Ph.D. dissertation, University of Massachusetts, Amherst.

[90] Heim, I. (1983), "On the projection problem for presuppositions", In: M. Barlow et al. (eds.), *Proceedings of the Second West Coast Conference on Formal Linguistics*, Stanford University, Standford, pp. 114-123.

[91] Heim, I. (1990), "E-type pronouns and donkey anaphora", *Linguistics and Philosophy*, 13, pp. 137-178.

[92] Heim, I. (1992), "Presupposition projection and the semantics of attitude verbs", *Journal of Semantics*, 9, pp. 183-221.

[93] Hintikka, J. (1962), *Knowledge and Belief*, Cornell University Press, Ithaca.

[94] Jeffrey, R. (1965), *The Logic of Decision*, University of Chicago Press, Chicago.

[95] Jeffrey, R. and R. Stalnaker (1994), "Conditionals as random variables", In: E. Eells and B. Skyrms (eds), *Probability and conditionals*, Cambridge University Press, Cambridge, pp. 31-46 .

[96] Kadmon, N. (1990), "Uniqueness", *Linguistics and Philosophy*, 13, pp. 273-324.

[97] Kálmán, L. and G. Rádai (1998), "Indefinites and even less definites", In: L. Kálmán et al (eds.), *Proceedings of the 6th Symposium on Logic and Language*, Budapest.

[98] Kamp, H. (1971), "Formal properties of 'Now' ", *Theoria*, 37, pp. 227-273.

[99] Kamp, H. (1981), "A theory of truth and semantic representation", In: J. Groenendijk et al. (eds), *Formal Methods in the Study of Language*, Amsterdam, pp. 277-322.

[100] Kamp, H. (1988), "Comments on Robert Stalnaker: 'Belief attribution and context' ", In: R. Grimm and D. Merrill (eds.), *Contents of Thought*, University of Arizona Press, Tuscon, pp. 16-181.

[101] Kamp, H. (1990), "Prolegomena to a structural account of belief and other attitudes", In: C. A. Anderson and J. Owens (eds.), *Propositional Attitudes, The Role of Content in Logic, Language, and Mind*, CSLI Lecture Notes, Nr. 20, Standford, pp. 27-90.

[102] Kamp, H. and U. Reyle, (1993), *From Discourse to Logic*, Kluwer, Dordrecht.

[103] Kaplan, D. (1969), "Quantifying in", In: D. Davidson and J. Hintikka (eds.), *Words and Objections, Essays on the work of W.V. Quine*, Reidel, Dordrecht, pp. 178-214.

[104] Kaplan, D. (1978), "Dthat", In: P. Cole (ed.), *Syntax and Semantics, Vol. 9: Pragmatics*, Academic Press, New York, pp. 221-243.

[105] Kaplan, D. (1989), "Demonstratives", In: I. Almog et al. (eds.), *Themes from Kaplan*, Oxford University Press, New York, pp. 481-563.

[106] Karttunen, L. (1969), "Pronouns and variables", In: R. I. Binnick et al. (eds), *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*, Chicago, pp. 108-115.

[107] Karttunen, L. (1974), "Presuppositions and linguistic context", *Theoretical Linguistics*, 1, pp. 181-194.

[108] Karttunen, L. and S. Peters, (1979), "Conventional implicature", In: C.K. Oh and D. Dinneen (eds.), *Syntax and Semantics, vol 11: Presupposition*, Academic Press, New York, pp.1-56.

[109] Katsuno, H. and A. Mendelzon, (1991), "On the difference between updating a knowledge base and revising it", In: P. Gärdenfors (ed.), *Belief Revision, Cambridge Tracts in Theoretical Computer Science, no. 29*, Cambridge Univesity Press, Cambridge, pp. 183-203.

[110] Kibble, R. (1994), "Dynamics of epistemic modality and anaphora", In: H. Bunt et al. (eds.), *Proceedings of the International Workshop on Computational Semantics*, Tilburg, pp. 121-130.

[111] Krahmer, E. and R. Muskens (1995), "Negation and disjunction in discourse representation theory", *Journal of Semantics*, 12, pp. 357-376.

[112] Kratzer, A. (1981), "Partition and revision: the semantics of counterfactuals", *Journal of Philosophical Logic*, 23, pp. 35-62.

[113] Kratzer, A. (1989), "An investigation of the lumps of thought", *Linguistics and Philosophy*, 12, pp. 607-653.

[114] Kratzer, A. (1998), "Scope or Pseudoscope? Are there wide scope indefinites?", In: S. Rothstein (ed.), *Events and Grammar*, Dordrecht, Kluwer, pp. 163-196.

[115] Kripke, S. (1971), "Identity and necessity", In: M. Munitz (ed.), *Identity and Individuation*, New York University Press, New York, pp. 135-164.

[116] Kripke, S. (1972), "Naming and necessity", In: D. Davidson and G. Harman (eds.), *Semantics of Natural Language*, Reidel, Dordrecht, pp. 253-355, 763-769. Revised and enlarged revision first published in 1980 by Blackwell, Oxford.

[117] Kripke, S. (1977), "Speakers reference and semantic reference", In: P. French et al. (eds.), *Studies in the Philosophy of Language*, Midwest Studies in Philosophy, no. 2, University of Minnesota Press, Minneapolis, pp. 255-276.

[118] Kripke, S. (1979), "A puzzle about belief", In: A. Margalit (ed.), *Meaning and Use*, Reidel, Dordrecht, pp. 239-283.

[119] Kripke, S. (ms), *Presupposition and anaphora: Remarks on the formulation of the projection problem*, Princeton University.

[120] Landman, F. (1986), "Conflicting presuppositions and modal subordination", In: *Papers from the 22nd Regional Meeting, Chicago Linguistic Society*, pp. 195-207.

[121] Lerner, J.Y. and T.E. Zimmermann, (1984), "Bedeutung und Inhalt von Eigennamen", *Papier Nr. 94 des SFB 99*, Konstanz.

[122] Lewis, D.K. (1968), "Counterpart theory and quantified modal logic", *The Journal of Philosophy*, 65, pp. 113-126.

[123] Lewis, D.K. (1969), *Convention*, Harvard University Press, Cambridge, Mass.

[124] Lewis, D.K. (1973), *Counterfactuals*, Blackwell, Oxford.

[125] Lewis, D.K. (1974), "Radical interpretation", *Synthese*, 23, pp. 331- 344.

[126] Lewis, D.K. (1975), "Probabilities of conditionals and conditional probabilities", *The Philosophical Review*, 85, pp. 297-315.

[127] Lewis, D.K. (1979a), "Attitudes de dicto and de se", *Philosophical Review*, 88, pp. 513-543.

[128] Lewis, D.K. (1979b), "Scorekeeping in a language game", *Journal of Philosophical Logic*, 8, pp. 339-359.

[129] Lewis, D.K. (1979c), "Counterfactual dependence and time's arrow", *Nous*, 13, pp. 455-476.

[130] Lewis, D.K. (1980), "Index, context and content", In: S. Kanger and S. Ohman (eds.), *Philosophy and Grammar*, Reidel, Dordrecht, pp. 79-100.

[131] Lewis, D.K. (1981), "What puzzling Pierre does not believe", *Australasian Journal of Philosophy*, 59, pp. 283-289.

[132] Lewis, D.K. (1982), "'Whether' reports", In: T. Pauli et al. (eds.), 320311: *Philosophical Essays Dedicated to Lennart Aqvist on his Fiftieth Birthday*, Filosofiska Studier, Upsala, pp. 194-206.

[133] Lewis, D.K. (1986), *On the Plurality of Worlds*, Basic Blackwell, Oxford.

[134] Lewis, D.K. (1996), 'Elusive Knowledge', *The Australian Journal of Philosophy*, 74, pp. 549-567.

[135] McGee, V. (1994), "Learning the impossible", In: E. Eells and B. Skyrms (eds), *Probability and Conditionals*, Cambridge University Press, Cambridge, pp. 179-199.

[136] McKay, T. and P. van Inwagen (1977), "Counterfactuals with disjunctive antecedents", *Philosophical Studies*, 31, pp. 353-356.

[137] Montague, R. (1974), *Formal Philosophy*, Yale University Press, New Haven.

[138] Morreau, M. (1992), *Conditionals in Philosophy and Artificial Intelligence*, Ph.D. dissertation, University of Stuttgart.

[139] Muskens, R. (1989), *Meaning and Partiality*, Ph.D. dissertation, University of Amsterdam.

[140] Neale, S. (1990), *Descriptions*, MIT Press, Cambridge.

[141] Nute, D. (1984), "Conditional logic", In: D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic, Vol II*, D. Reidel, Dordrecht, pp. 387-440.

[142] Partee, B. (1972), "Opacity, coreference, and pronouns", In: D. Davidson and G. Harman (eds.), *Semantics of Natural Language*, Reidel, Dordrecht, pp. 415-441.

[143] Peregrin, J. and K. von Heusinger (1997), "Dynamic semantics with choice functions", In: H. Kamp & B. Partee (eds.), *Proceedings of the Workshop "Context Dependence in the Analysis of Linguistic Meaning*, Stuttgart/Prague, pp. 329-354.

[144] Perry, J, (1977), "Frege on demonstratives", *Philosophical Review*, 86, pp. 474-497.

[145] Perry, J. (1979), "The problem of the essential indexical", *Nous*, 13, pp. 3-31.

[146] Peters, S. (1977), "A truthconditional formulation of Karttunen's account of presuppositions", In: *Texas Linguistic Forum 6*, Department of Linguistics, University of Texas at Austin, pp. 137-149.

[147] Popper, K.R. (1959), *The Logic of Scientific Discovery*, Hutchinson, London.

[148] Portner, P. (1997), "The semantics of mood, complementation, and conversational force", *Natural Language Semantics*, 5, pp. 167-212.

[149] Powers, L. (1976), "Comments on 'Propositions' ", In: A. Mackay and D. Merrill (eds.), *Issues in the Philosophy of Language*, Yale University Press, New Haven, pp. 93-103.

[150] Price, H. (1989), "Defending desire-as-belief", *Mind*, 88, pp. 119-127.

[151] Putnam, H. (1975), "The meaning of 'meaning' ", In: K. Gunderson (ed.), *Language, Mind and Knowledge*, University of Minnesota Press, , Minneapolis, pp. 131-193.

[152] Quine, W.V. (1953), "Reference and modality", In: W.V. Quine, *From a Logical Point of View*, Harvard University Press, Cambridge.

[153] Quine, W.V. (1956), "Quantifiers and propositional attitudes", *The Journal of Philosophy*, 53, pp. 177-187.

[154] Quine, W.V. (1960), *Word and Object*, MIT Press, Cambridge Mass.

[155] Ramsey, F.P. (1931), *The Foundations of Mathematics and other Logical Essays*, R.B. Braithwaite (ed.), Harcourt Brace, New York.

[156] Rescher, N. (1967), "Semantic foundations for the logic of preference", In: N. Rescher (ed.), *The Logic of Decision and Action*, University of Pittsburgh Press, Pittsburgh.

[157] Richard, M. (1983), "Direct reference and ascription of belief", *Journal of Philosophical Logic*, 12, pp. 425-452.

[158] Roberts, C. (1989), "Modal subordination and pronominal anaphora in discourse", *Linguistics and Philosophy*, 12, pp. 683-721.

[159] Rooy, R. van (1997), *Attitudes and Changing Contexts*, Ph.D. dissertation, University of Stuttgart.

[160] Rooy, R. van (1998), "Modal subordination in questions", In: J. Hulstijn and A. Nijhold, (eds.), *The Proceedings of Twendial '98*, Enschede, pp. 237-248.

[161] Rooy, van R. (2000), "Anaphoric relations across belief contexts", In: K. von Heusinger and U. Egli (eds.), *Reference and Anaphoric Relations*, Kluwer, Dordrecht, pp. 157-182.

[162] Rooy, van R. (2001), "Exhaustivity in dynamic semantics; Referential and Descriptive pronouns", *Linguistics and Philosophy*, 24, pp. 621-657.

[163] Rooy, van R. (to appear), "A modal analysis of presupposition and modal subordination", *Journal of Semantics*.

[164] Russell, B. (1905), "On denoting", *Mind*, 14, pp. 479-493.

[165] Russell, B. (1948), *Human Knowledge: Its Scope and Limits*, George Allen & Unwin Ltd., London.

[166] Saarinen, E. (1978), "Intentional identity interpreted", *Linguistics and Philosophy*, 2, pp. 151-223.

[167] Salmon, N. (1986), *Frege's Puzzle*, MIT Press, Cambridge.

[168] Sandt, R.A. van der (1982), *Kontekst en Presuppositie*, Ph.D. Dissertation, University of Nijmegen.

[169] Sandt, R.A. van der (1988), *Context and Presupposition*, Croom Helm, London.

[170] Sandt, R.A. van der (1992), "Presupposition projection as anaphora resolution", *Journal of Semantics*, 9, pp. 223-267.

[171] Segerberg, K. (1973), "Two dimensional modal logic", *Journal of Philosophical Logic*, 2, pp. 77-96.

[172] Shackle, G.l.S. (1961), *Decision, Order and Time in Human Affairs*, Cambridge University Press, Cambridge.

[173] Skyrms, B. (1980a), *Causal Necessity*, Yale University Press, New Haven, Conn.

[174] Skyrms, B. (1980b), "The prior prospensity account of subjunctive conditionals", In: W.L. Harper et al. (eds), *Ifs*, Reidel, Dordrecht, pp.259-265.

[175] Skyrms, B. (1994), "Adams conditionals", In: E. Eells and B. Skyrms (eds), *Probability and conditionals*, Cambridge University Press, Cambridge, pp. 13-26.

[176] Slater, B.H. (1988), "Intensional identities", *Logique et Analyse*, 2, pp. 93-107.

[177] Soames, S. (1982), "How presuppositions are inherited: a solution to the projection problem", *Linguistic Inquiry*, 13, pp. 483-545.

[178] Sommers, F. (1982), *The Logic of Natural Language*, Clarendon Press, Oxford.

[179] Spohn, W. (1987), "Ordinal conditional functions: a dynamic theory of epistemic states", W.L. Harper and B. Skyrms (eds.), In: *Causation in Decision, Belief Change, and Statistics, vol. 2*, Reidel, Dordrecht, pp. 105-134.

[180] Spohn, W. (1997), *Begründungen a prioi - oder: ein frischer Blick auf Dispositionspredikate*, In: W. Lenzen (ed.), *Des weite Spectrum der Analytischen Philosophie. Festschrifft for Franz von Kutschera* de Gruyter, Berlin, pp. 323-345.

[181] Stalnaker, R.C. (1968), "A theory of conditionals", *Studies in Logical Theory, American Philosophical Quarterly Monograph Series, No. 2*, Blackwell, Oxford, pp. 98-112.

[182] Stalnaker, R.C. (1970a), "Probability and conditionals", *Philosophy of Science*, 37, 64-80.

[183] Stalnaker, R.C. (1970b), "Pragmatics", *Synthese*, 22, pp. 272-289.

[184] Stalnaker, R.C. (1972), "Propositions", In: A. Mackay and D. Merrill (eds.), *Issues in the Philosophy of Language*, Yale University Press, New Haven and London, pp. 197- 213.

[185] Stalnaker, R.C. (1973), "Presuppositions", *Journal of Philosophical Logic*, 2, pp. 447-457.

[186] Stalnaker, R.C. (1974), "Pragmatic presuppositions", In: M. K. Munitz and P. Unger (eds.), *Semantics and Philosophy*, New York Univerisity Press, New York, pp. 197-213.

[187] Stalnaker, R.C. (1975), "Indicative conditionals", *Philosophia*, 5, pp. 269-286.

[188] Stalnaker, R.C. (1976a), "Letter to Bas van Fraassen", In: W.L. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. I*, Reidel, Dordrecht, pp. 302-306.

[189] Stalnaker, R.C. (1976b), "Letter to William Harper", In: W.L. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. I*, Reidel, Dordrecht, pp. 113-115.

[190] Stalnaker, R.C. (1977), "Complex predicates", *The Monist*, 60, pp. 327-339.

[191] Stalnaker, R.C. (1978), "Assertion", In: P. Cole (ed.), *Syntax and Semantics, vol. 9: Pragmatics*, Academic Press, New York, pp. 315-332.

[192] Stalnaker, R.C. (1979), "Anti-essentialism", In: P. French et al. (eds.), *Midwest Studies in Philosophy*, 4, pp. 343-355.

[193] Stalnaker, R.C. (1980a), "A defense of conditional excluded middle", In: *Ifs*, W.L. Harper et al. (eds.), Reidel, Dordrecht, pp. 87-104.

[194] Stalnaker, R.C. (1980b), "Letter to David Lewis", In: *Ifs*, W.L. Harper et al. (eds.), Reidel, Dordrecht, pp. 151-152.

[195] Stalnaker, R.C. (1981), "Indexical belief", *Synthese*, 49, pp. 129-151.

[196] Stalnaker, R.C. (1984), *Inquiry*, MIT Press, Cambridge, Mass..

[197] Stalnaker, R.C. (1986), "Counterparts and identity", *Midwest Studies in Philosophy*, 11, pp. 121-140.

[198] Stalnaker, R.C. (1987), "Semantics for belief", *Philosophical Topics*, 15, pp. 177-190.

[199] Stalnaker, R.C. (1988), "Belief attribution and context", In: R. Grimm and D. Merrill (eds.) , *Contents of Thought*, University of Arizona Press, Tuscon, 156-181.

[200] Stalnaker, R.C. (1989), "On what's in the head", In: J.E. Tomberlin (ed.), *Philosophical Perspectives, 3: Philosophy of Mind and Action Theory*, Atascadero, Ridgeview, pp. 129-146.

[201] Stalnaker, R.C. (1990a), "Mental content and linguistic form", *Philosophical Studies*, 58, pp. 129-146.

[202] Stalnaker, R.C. (1990b), "Narrow content", In: C.A. Anderson and J. Owens (eds.), *Propositional Attitudes: The Role of Content in Logic, Language and Mind*, CSLI, Standford.

[203] Stalnaker, R.C. (1991), "The problem of logical omniscience, I", *Synthese*, 89, pp. 425- 440.

[204] Stalnaker, R.C. (1993), "Twin Earth revisited", *Proceedings of the Aristotelian Society*, 93, pp. 297-311.

[205] Stalnaker, R.C. (1994), "Stalnaker", In: S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*, Blackwell, Oxford, pp. 561-568.

[206] Stalnaker, R. C. (1996), "Knowledge, belief and counterfactual reasoning in games", *Economics and Philosophy*, 12, pp. 133-163.

[207] Stalnaker, R. C. (1998a), "Reference and necessity", In: B. Hale and C. Wright (eds.), *A Companion to the Philosophy of Language*, Blackwell, Oxford, pp. 534-554.

[208] Stalnaker, R.C. (1998b), "On the representation of context", *Journal of Logic, Language and Information*, 7, pp. 3-19.

[209] Stalnaker, R.C. (1999), *Context and Content*, Oxford University Press, Oxford.

[210] Stalnaker, R.C. (2001), "On considering a possible world as actual", *Proceedings of the Aristotelian Society*, 75, pp. 141-156.

[211] Stalnaker, R.C. (2002), "Common ground", *Linguistics and Philosophy*, 25, pp. 701-721.

[212] Stampe, D. W. (1977), "Towards a causal theory of linguistic representation", *Midwest Studies in Philosophy, II: Studies in the Philosophy of Language*, University of Minnesota at Morris, Morris, pp. 42-63.

[213] Stechow, A. von (1984), "Structured propositions and essential indexicals", In: F. Landman and F. Veltman (eds.), *Varieties of Formal Semantics*, Foris Publications, Dordrecht, pp. 385- 403.

[214] Stine, G. (1976), "Scepticism, relevant alternatives, and deductive closure", *Philosophical Studies*, 29, pp. 249-261.

[215] Strawson, P.F. (1950), "On referring", *Mind*, 59, pp. 320-344.

[216] Thijsse, E. (1992), *Partial Logic and Knowledge Representation*, Eburon Publishers, Delft.

[217] Thomason, R.H. and A. Gupta, (1980), "A theory of conditionals in the context of branching time", In: W.L. Harper et al. (eds.), *Ifs*, Reidel, Dordrecht, pp. 299-322.

[218] Veltman, F. (1996), "Defaults in update semantics", *Journal of Philosophical Logic*, 25, pp. 221-261.

[219] Warmbrod, K, (1981), "Counterfactuals and substitution of equivalent antecedents", *Journal of Philosophical Logic*, 10, pp. 267-289.

[220] Westerståhl, D. (1984), "Determiners and contexts sets", In: J. van Benthem and A ter Meulen (eds.), *Generalised Quantifiers in Natural Language*, Foris Publishers, Dordrecht, pp. 45-71.

[221] Wright, H. von (1963), *The Logic of Preference*, University Press, Edinburgh.

[222] Zeevat, H. (1996), "A neoclassical analysis of belief sentences", In: P. Dekker and M. Stokhof (eds.), *Proceedings of the 10th Amsterdam Colloquium*, Amsterdam, pp. 723-741.

[223] Zeevant, H. (1997), "The common ground as a dialogue parameter", In: A. Benz & G. Jäger, *Proceedings of Mundial'97*, CIS, pp. 195-214.

[224] Zimmermann, T.E. (1991), "Kontextabhangigkeit", In: A. von Stechow and D. Wunderlich (eds.), *Semantik: Ein Internationales Handbuch der Zeitgenossischen Forschung*, De Gruyter, Berlin, pp. 156-229.

[225] Zimmermann, T.E. (1997), "Remarks on the epistemic role of discourse referents", In: L. Moss, J. Ginzburg, M. de Rijke (eds.), *Logic, Language, and Computation. Vol. 2*, CSLI, Stanford.

[226] Zimmermann, T.E. (1999), "Scepticism de se", *Erkenntnis*, 51, pp. 267-275.