Interpreting Linguistic Behavior with Possible World Models

Johannes Marti

Interpreting Linguistic Behavior with Possible World Models

ILLC Dissertation Series DS-2016-04

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Printed and bound by GVO drukkers & vormgevers B.V.

ISBN: 978-94-6332-026-9

Interpreting Linguistic Behavior with Possible World Models

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 7 juni 2016, te 16.00 uur

door

Johannes Friedrich Marti

geboren te Bazel, Zwitserland.

Promotiecommissie

Promotors:	Prof. dr. M.J.B. Stokhof	Universiteit van Amsterdam
	Prof. dr. F.J.M.M. Veltman	Universiteit van Amsterdam
Overige leden:	Prof. dr. R.A.M. van Rooij	Universiteit van Amsterdam
	Prof. dr. F. Berto	Universiteit van Amsterdam
	Dr. A. Baltag	Universiteit van Amsterdam
	Prof. dr. A. Visser	Universiteit Utrecht
	Prof. dr. H.D. Rott	Universität Regensburg
	Dr. S. Yalcin	University of California, Berkeley

Faculteit der Geesteswetenschappen

Contents

Acknowledgments

First of all, I thank my supervisors, Martin Stokhof and Frank Veltman. Martin's apt, but unimposing, guidance helped me to write the thesis that I wanted to write. I especially remember the engaging discussions with Martin during which he made multiple remarks that I initially dismissed but then later noticed that they were spot on. From Frank I felt a cheerful curiosity about the technical parts of my work, which motivated me to keep going. I am also grateful for his effort to carefully read through the almost completed manuscript of the thesis and untangling some conceptual confusions.

I also thank Robert van Rooij, Franz Berto, Alexandru Baltag, Albert Visser, Hans Rott and Seth Yalcin, for agreeing to be in the committee. Additionally, I thank: Robert, for keeping asking me about my thesis topic, even at times when I had nothing to reply, and for pointing out some crucial references; Alexandru, for encouraging feedback after a seminar talk that motivated me at the beginning of the writing process; Albert, for sending me a list of corrections that fixed some of the worst spelling mistakes in the manuscript; And Hans, for discussions about verbal disagreement and belief revision, two topics that are closely related to the topic of this thesis.

This thesis profited a lot from my interaction with colleagues at the ILLC, some of which I want to mention explicitly. I am indebted to Peter Fritz for patiently listening to my ideas when I was applying for the PhD position, and especially, for some encouraging words after one and a half years in the thesis project, when I had no idea what I was doing and my motivation had completely collapsed. One of the best parts of my PhD experience was working together with Riccardo Pinosio on conditional logic and non-monotonic reasoning. Some of the results from this collaboration are crucial for the proofs in Section 7.4 of this thesis. Lastly, I need to mention Paolo Galeazzi who joined the ILLC and, coincidentally, became my flatmate, just right at the time when I needed someone to help me with decision theory. The conversations I had with Paolo about representation results in decision theory and the notion of rationality strongly influenced my conceptual understanding of the results in this thesis.

I also want to thank Matthijs Westera for translating my abstract to Dutch and Corina Koolen for proofreading the translated abstract in the last moments before the manuscript was submitted to the committee.

The thesis greatly profited from the inspiring interdisciplinary atmosphere at the ILLC. I could probably not have written this thesis anywhere else. I am incredibly thankful to all the people that make the ILLC the place it is, and this explicitly includes the administrative staff and the management.

I also owe a lot of gratitude to all the people that were behind the philosophy of language group in Berne at the time when I was a Bachelor student there. There I obtained most of the philosophical knowledge that was necessary for writing this thesis. I explicitly remember Jonas Pfister, in whose proseminars I learned about Davidson for the first time, and Sebastian Leugger, who was organizing wonderful reading groups about two-dimensional semantics, Carnap and metaethics.

On the less academic side I want to thank all the people that made me have a good time in Amsterdam. Especially, I thank: Sumit Sourabh, for many hours of playing table tennis; Ivano Ciardelli, for good food and conversations about logic; And I thank Vlasta Sikimić for cheering me up uncountable times when I was in a bad mood.

Lastly, I grateful to my sister, Annegret, for being an amazing sister, and my mother, Hella, for her unconditional support. I dedicate this book to my mother.

Amsterdam   Johannes Marti
April, 2016.

Chapter 1

Introduction

This thesis is about the problem of interpreting linguistic behavior with possible world models for belief and meaning. The framework of possible worlds has been applied in doxastic logic to represent the beliefs of subjects and in formal semantics to represent the meaning of sentences in natural languages. In both settings the central notion that is investigated, belief and meaning respectively, is itself a theoretical notion. The aim of this thesis is to connect these notions of belief and meaning with the more empirical notion of linguistic behavior by giving a formal account of how possible world models for belief and meaning can be used to interpret linguistic behavior.

The setting of this thesis is as follows: We imagine that we are observing some subject and want to find a possible world model that represents the beliefs of the subject and the meaning of sentences in the language that the subject is speaking. To this aim we interpret the linguistic behavior of the subject. We construct a model that ascribes such beliefs to the subject and such meanings to the sentences in her language so that we can make sense of her linguistic behavior as resulting from these beliefs and meanings.

Interpreting the linguistic behavior of some subject with a possible world model requires that we have a formal account of how the representation of beliefs and meanings in a possible world model relates to the linguistic behavior of the subject. Giving such a formal account allows us to prove representation results, which show that if and only if the linguistic behavior of the subject meets certain conditions then it can be taken to arise from a possible world model in a certain class of models.

In this thesis I consider many distinct classes of possible world models that represent beliefs and meanings with different mathematical structures. I am interested in comparing different kinds of formal models and investigating what conditions they impose on linguistic behaviors rather than arguing for one particular class of models or defending some particular conditions on linguistic behaviors

In the following sections of this chapter, I give an introduction to the possible world framework, explain the problem of interpreting linguistic behavior and sketch the approach that I take later in the thesis to solve this problem.

Section 1.1 introduces possible world models for belief, which represent the beliefs of some subject using the possible world framework. In Section 1.2 I explain the idea of possible world semantics, which is to represent the meaning of sentences with sets of possible worlds.

In Section 1.3 I describe the problem of interpreting linguistic behavior with possible world models and explain that it is an instance of the problem of radical interpretation that has been discussed in the philosophical literature.

In Section 1.4 I sketch the common structure of the accounts of interpretation that I give later in the thesis. I also set up three criteria with which I asses and compare the different accounts of interpretation developed in this thesis.

Section 1.5 provides an overview of the structure of this thesis.

1.1 Possible world models for belief

In this section I explain what possible world models are and how the beliefs of some subject are represented in a possible world model.

Possible world models are based on a set W which is the *domain* of the model. The elements of the domain are called *possible worlds* or just *worlds*. Intuitively, one thinks of a world $w \in W$ as a way how things might be. It is assumed that for any world and any fact that is relevant in the application at hand it is specified whether the fact obtains at the world. One way to ensure this is to describe for every world in the domain what relevant facts obtain at that world. I am using the term *basic fact* for those facts that are relevant for the specific application and hence are specified to either obtain or not obtain at every world in W. It is also assumed that the domain contains at least one possible world for every combination of basic facts that is relevant for the application at hand.

Let us consider an example in which it is relevant whether it is raining and whether there are raindrops on the window. We use a domain $W = \{w, v, u\}$ containing three possible worlds: In world w it is raining and there are raindrops on the window. In v it is not raining but there are still raindrops on the window. In u it is also not raining and the raindrops on the window have dried away. The domain W does not include a world in which it is raining but there are no raindrops on the window. We assume that for us this is just not a relevant way how things might be.

One possible world corresponds to the way how things really are in the situation in which we are interpreting the subject. This special world is called the *actual world*. The basic facts that are specified to obtain at the actual world are those basic facts that indeed obtain in the situation in which we are interpreting the subject. It is presupposed that this is always a relevant combination of basic facts that is included in the domain. As an example imagine that the subject is in a situation where it is not raining but there are still some raindrops on the window. In this case the actual world would be the element v from the domain in the example above.

A *proposition* is a subset $P \subseteq W$ of the domain of all possible worlds. If there is need to make the domain W explicit relative to which a proposition $P \subseteq W$ is defined I also say that P is a *proposition over* W. I use the phrase the proposition that such and such to denote the set of all worlds where the fact that such and such obtains. A proposition $P \subseteq W$ *true* at a world $w \in W$ if $w \in P$ and it is *false* at w if $w \notin P$. In our example above we have for instance that $\{w, v\} \subseteq W$ is the proposition that there are raindrops on the window, which is true at w and v and false at u.

A proposition $Q \subseteq W$ *implies* a proposition $P \subseteq W$ if $Q \subseteq P$. Similarly, a set of propositions $\mathcal{U} \subseteq \mathcal{P}W$ *implies* a proposition $P \subseteq W$ if $\bigcap \mathcal{U} \subseteq P$. In the example above we have for instance that the proposition that it is raining implies the proposition that there are raindrops on the window because $\{w\} \subseteq \{w, v\}$. Relative to a domain it is possible that some proposition implies another even though there is no clear sense in which the latter would be a logical consequence of the former. That some proposition implies another is not a matter of syntactic consequence in some formal system but just depends on which possible worlds we choose to include in the domain.

Relative to a domain W one can represent the belief state of some subject by a subset $B \subseteq W$. This set B is called the *belief set* of the subject and the elements of B are called *doxastic alternatives* of the subject. Given a possible world $w \in W$ one also says that the subject *considers the world w possible* if w is a doxastic alternative of the subject, that is, $w \in B$. Intuitively, a world w is a doxastic alternative for the subject if it is compatible with everything that the subject believes.

A *possible world model for belief* is a pair (W, B) where W is a domain of possible worlds and $B \subseteq W$ is the belief set of some subject. According to a possible world model (W, B) the subject whose beliefs are represented in the model *believes a proposition P* if $B \subseteq P$. To express the same thing differently one could also say that the subject believes the proposition P if P is true at all of her doxastic alternatives. I also say that according to some model the subject believes that such and such is the case whenever she believes the proposition that is the set of all worlds where such and such is true. Similarly, the belief that such and such is the proposition that contains all worlds where such and such is true, presupposing that this proposition is believed by the subject in a given model. According to the model (W, B) the subject *considers a proposition P possible* if $P \cap B \neq \emptyset$. To express this differently: The subject considers the proposition P possible if P is true at least one of her doxastic alternatives. Again, I say that the subject considers such and such possible whenever she considers the proposition that is the set of worlds where such and such is true possible.

Let us have a look at some examples of possible world models that represent the beliefs of some subject about the weather and the presence of raindrops on some window. In all these models we again use the domain $W = \{w, v, u\}$ from the example above.

First consider the model $(W, \{w\})$. In this model w is the only doxastic alternative of the subject. According to the model the subject believes for instance that it is raining, because it is raining in all of her doxastic alternatives. Similarly, she also believes that there are raindrops on the window. She does not believe the raindrops have dried up and she does not believe that the sun is shining because both of these propositions are false at one of her doxastic alternatives. The model $(W, \{w\})$ does not give us any information to whether the subject believes that, say, she has an umbrella in her bag or whether she believes that, say, the continuum hypothesis is true. If we were interested in representing her beliefs about these propositions we would need a different set of worlds as the domain of our model.

Let us consider another model $(W, \{v, u\})$. This model represents a situation where the subject believes that it is not raining but she is uncertain whether there are raindrops on the window. She does not believe that there are raindrops nor that there are none. To put this differently, she considers it possible that there are no raindrops on the window and she considers it possible that there are raindrops on the window.

An extreme case is the model (W, W) where all worlds are doxastic alternatives for the subject. In this model the subject has no beliefs about any particular facts. She considers it possible that it is raining and she considers it possible that it is not raining. Similarly, for the raindrops being on the window. However, she does for instance believe that if it is raining then there are raindrops on the window, since in all her doxastic alternatives where it is raining there are indeed raindrops on the window.

Another extreme case is the model $(W, \emptyset)$ where the belief set of the subject is the empty set $\emptyset$. In this model the subject believes every proposition and considers no proposition possible. We would usually not expect the subject to be in this state.

A consequence of using simply a set of possible worlds to represent all of the subject's beliefs is that beliefs are closed under implication of propositions. The subject believes a proposition P if $B \subseteq P$ where P is her belief set. Now for any other proposition Q which is implied by P, meaning that $P \subseteq Q$, it follows that $B \subseteq Q$ and so the subject also believes Q. The assumption that beliefs are closed under implication allows us to represent a set of beliefs by just a proposition instead of a set of propositions. Given a set of beliefs $\mathcal{U} \subseteq \mathcal{P}W$ we can take the intersection $\bigcap \mathcal{U} \subseteq W$ of all beliefs in $\mathcal{U}$. This is again a proposition and we have that some subject believes $\bigcap \mathcal{U}$ if and only if she believes every proposition that is implied by the propositions in $\mathcal{U}$. The belief set B of some subject can be thought of as the intersection $\bigcap \mathcal{B}$ over the set $\mathcal{B} \subseteq \mathcal{P}W$ of all propositions that the subject believes.

The possible worlds models introduced here, which are just a domain of possible worlds plus a belief set for the subject, are the simplest kind of possible world models that I am considering. Later in the thesis I introduce various extensions of these basic models.

References to the literature

Possible world models for belief are mostly studied in the context of doxastic logics. Doxastic logics are modal logics in which the modality expresses that some subject believes that something is the case. Doxastic logics are often treated as a simple modification of epistemic logics, in which the modality expresses that some subject knows that something is the case. Hintikka (1962) was the first to use a semantics akin to possible world semantics for epistemic and doxastic logic. For a recent account of possible world models for epistemic and doxastic logics I refer to Fagin et al. (2003) and to van Ditmarsch, van der Hoek, and Kooi (2007).

In doxastic logic it is common to not only consider the beliefs that a subject has about the basic facts but also the beliefs that she has about her own beliefs and the beliefs of others. Much of the complexity in doxastic logic results from representing this iterative nature of belief. I do not consider such higher-order beliefs in this thesis.

A lot of the philosophical literature is concerned with the metaphysical and epistemological status of possible worlds (see Menzel 2015 for an overview and references). I want to bother with these questions as little as possible and take possible worlds to be just a mathematical tool for modeling linguistic behavior. There are however two peculiarities of my use of possible worlds that I want to mention explicitly because they depart from assumptions that are commonly made in the literature.

First, I take possible worlds to describe only a relatively small portion of reality that might depend on the subject's location in time and space. For instance the fact that it is raining might be true at some times and places and false at others. This is different from the common view that possible worlds determine a whole temporal and spatial reality. On this view there would be no fact that it is raining. One can only say that it is raining on some particular time at some particular place. It might be possible to simulate possible worlds as I use them in this thesis with tuples that contain a possible world that is a complete temporal and spacial reality together with a point in time and space of that reality. Such tuples have been called centered worlds by Lewis (1979). Lewis suggests using a set of centered worlds as the belief set of the subject to account for the beliefs that the subject has about her own position in the world.

Second, I assume that there is more than one single totality of possible worlds. In every modeling context we might use a different domain of possible worlds, depending on what basic facts and combinations thereof are relevant. This means that the domain of possible worlds can change in at least two ways. The domain increases or shrinks if we start or stop considering some combinations of basic facts as relevant. We might also start considering some additional facts as relevant, in which case we get a splitting of the existing possible worlds, or we might start neglecting some previously relevant facts, in which case we get a new domain of worlds onto which the previous domain projects.

1.2 Possible world semantics

In this section I show how to extend the notion of a possible world model for belief to include a representation of the meaning of sentences in the language of the subject. I first discuss the formal representation of the sentences which the subject uses and then describe how to assign meanings to these sentences.

Throughout this thesis I am using the symbol $\mathcal{V}$ for the set of all *sentences* that the subject is using. I call the set $\mathcal{V}$ the *vocabulary* of the subject. The sentences in the vocabulary are uninterpreted expressions. They are types of similar sign tokens such as sequences of sounds or strings of symbols in some alphabet. If for instance the subject is a speaker of English then $\mathcal{V}$ could contain the sentences "It is raining." and "There are no raindrops on the window." But it might also be that the subject uses completely different sentences in which case $\mathcal{V}$ could contain for instance the expressions "Pui pui." or "Ling ne ling."

In the simplest case we do not represent any syntactic structure in the sentences of the subject. In this case we take $\mathcal{V} = \mathsf{At}$, where At is any set. The elements of At are called *atomic sentences* and referred to by letters such as $p, q, r, \ldots$. For instance the atomic sentence p might stand for "It is raining.", q for the sentence "There are raindrops on the window." and r for "It is not raining but there are raindrops on the window." Note that the latter sentence is represented by the atomic sentence r, even though in English it has a rich syntactic structure involving p and q as its parts.

I also consider the case where we have a hypothesis about which syntactic constructions in the language of the subject function as the propositional connectives of classical logic. In this case we take $\mathcal{V} = \mathcal{B}$, where $\mathcal{B}$ is the set of propositional formulas over some fixed set At of atomic sentences. As an example consider again the sentence "It is not raining but there are raindrops on the window." If we suppose that the "but" in this sentence expresses a classical conjunction, and the "not" expresses a classical negation then we can represent this sentence with propositional formula $\neg p \wedge q$, where p and q are as above.

To interpret the sentences in the vocabulary of the subject we have to assign meanings to these sentences. Formally, this is done by a function from the set of sentences to a set of meanings. In standard possible world semantics the meaning of a sentence is a proposition, that is, a set of possible worlds. Intuitively, one thinks of the meaning of a sentence as the set of worlds where the sentence is true. So we can determine the meaning of sentences in $\mathcal{V}$ with a function $I : \mathcal{V} \to \mathcal{P}W$, that maps sentences to subsets of the domain W of possible worlds. Such functions $I : \mathcal{V} \to \mathcal{P}W$ are called *interpretation functions* or just *interpretations*. Given an interpretation function $I : \mathcal{V} \to \mathcal{P}W$ and a sentence $\varphi \in \mathcal{V}$ we call the set of worlds $I(\varphi)$ the *proposition expressed by* φ. If one has fixed an interpretation function then one can apply to sentences the same terminology as the one introduced in Section 1.1 for the propositions expressed by these sentences. For instance we can say that *a sentence* φ *is true at* a world w if $w \in I(\varphi)$. Or, relative to an interpretation I, ψ *implies* φ if $I(\psi) \subseteq I(\varphi)$.

If $\mathcal{V} = \mathsf{At}$ then an interpretation function $I : \mathsf{At} \to \mathcal{P}W$ is just a function that maps atomic sentence to propositions. As an example consider the English sentences "It is raining.", represented by the letter p, and "There are raindrops on the window.", represented by q. We assign the meaning of these sentences relative to the domain $W = \{w, v, u\}$ where w is the only world in which it is raining and in w and v but not u there are raindrops on the window. The meaning of p and q in English is then captured by an interpretation function $I : \mathsf{At} \to \mathcal{P}W$ such that $I(p) = \{w\}$ and $I(q) = \{w, v\}$.

In the case where $\mathcal{V} = \mathcal{B}$, we are making the assumption that the propositional connectives in the language of the subject have a classical semantics. This assumption is captured by considering only those interpretation functions $I : \mathcal{B} \to \mathcal{P}W$ that satisfy the following semantic clauses:

$$\begin{aligned}
I(\varphi \wedge \psi) &= I(\varphi) \cap I(\psi) = \{w \in W \mid w \in I(\varphi) \text{ and } w \in I(\psi)\}\\
I(\varphi \vee \psi) &= I(\varphi) \cup I(\psi) = \{w \in W \mid w \in I(\varphi) \text{ or } w \in I(\psi)\}\\
I(\neg\varphi) &= W \setminus I(\varphi) = \{w \in W \mid \text{not } w \in I(\varphi)\}\\
I(\varphi \to \psi) &= (W \setminus I(\varphi)) \cup I(\psi) = \{w \in W \mid \text{if } w \in I(\varphi) \text{ then } w \in I(\psi)\}\\
I(\bot) &= \emptyset.
\end{aligned}$$

For any interpretation $I : \mathsf{At} \to \mathcal{P}W$ there is a unique interpretation $I' : \mathcal{B} \to \mathcal{P}W$ satisfying the above clauses such that $I'(p) = I(p)$ for all $p \in \mathsf{At}$. Moreover any interpretation of type $I' : \mathcal{B} \to \mathcal{P}W$ that satisfies the above clauses is determined in this way by a unique interpretation $I : \mathsf{At} \to \mathcal{P}W$. For this reason it is sufficient to consider interpretation functions $I : \mathsf{At} \to \mathcal{P}W$ which map atomic sentences to propositions even if we are working under the assumption that $\mathcal{V} = \mathcal{B}$. Given an interpretation $I : \mathsf{At} \to \mathcal{P}W$ I am then also using the symbol I for its unique extension $I' : \mathcal{B} \to \mathcal{P}W$ to a function from propositional formulas to propositions satisfying the above clauses.

Consider again the interpretation $I : \mathsf{At} \to \mathcal{P}W$ from the example above. It is defined such that $I(p) = \{w\}$ and $I(q) = \{w, v\}$, which represents the meanings of "It is raining." and "There are raindrops on the window." in English. If we extend this to an interpretation $I : \mathcal{B} \to \mathcal{P}W$ on propositional formulas we have for instance that the sentence "It is not raining but there are raindrops on the window." expresses the proposition $I(\neg p \wedge q) = (W \setminus I(p)) \cap I(q) = \{v\}$.

One can also define a notion of logical consequence between sentences in $\mathcal{B}$. A sentence $\varphi \in \mathcal{B}$ is a *logical consequence* of a set of sentences $\Sigma \subseteq \mathcal{B}$ if for every set of worlds W and interpretation $I : \mathsf{At} \to \mathcal{P}W$ it holds that $\bigcap\{I(\psi) \mid \psi \in \Sigma\} \subseteq I(\varphi)$. This definition only quantifies over interpretations $I : \mathsf{At} \to \mathcal{P}W$. Equivalently, we could also quantify over all interpretations $I : \mathcal{B} \to \mathcal{P}W$ that satisfy the above semantic clauses for the propositional connectives. We also write $\Sigma \models \varphi$ to express that φ is a logical consequence of Σ. If $\Sigma = \{\psi_1, \dots, \psi_n\}$ is finite or empty then we also write that $\psi_1, \dots, \psi_n \models \varphi$ or that $\models \varphi$. In the latter case φ is also called a *tautology*. We write $\mathsf{cl}\,(\Sigma) \subseteq \mathcal{B}$ for the set of all logical consequences of $\Sigma \subseteq \mathcal{B}$, that is,

$$\mathsf{cl}\,(\Sigma) = \{\varphi \in \mathcal{B} \mid \Sigma \models \varphi\}.$$

The set $\mathsf{cl}\,(\Sigma)$ is called the *logical closure of* $\Sigma \subseteq \mathcal{B}$.

Note the difference between a sentence being a logical consequence of a set of sentences and a sentence being implied by a set of sentences. Implications between sentences are always relative to the assignment of meanings to sentences that is given by an interpretation function. The notion of logical consequence between sentences is independent from the meaning of those sentences. A sentence is a logical consequence of a set of sentences if it is implied by that set of sentences relative to every interpretation function. Hence, whenever a sentence is a logical consequence of a set of sentences then it is implied by it relative to every given interpretation function. The converse does not hold, as one can see on the example interpretation I from above with $I(p) = \{w\}$ and $I(q) = \{w, v\}$. Relative to this interpretation I the sentence p implies the sentence q. But q is not a logical consequence of p. A counterexample would for instance be the interpretation I' with $I'(p) = \{w, v\}$ and $I'(q) = \{v\}$ which swaps the meanings of p and q.

An important notion for this thesis is that of a set of sentences being closed under logical consequence. Hence we call a set of sentences $\Sigma \subseteq \mathcal{B}$ a *theory* if $\mathsf{cl}\,(\Sigma) \subseteq \Sigma$. A theory Σ is *consistent* if $\Sigma \neq \mathcal{B}$. A consistent theory Σ is *complete* if for every sentence $\varphi \in \mathcal{B}$ either $\varphi \in \Sigma$ or $\neg\varphi \in \Sigma$.

We now adapt the definition of a possible world model for belief to also contain information about the meaning of the sentences in the subject's language. To do so we add an interpretation function for the set At of atomic sentences in the language of the subject. A *simple possible world model for belief and meaning* or shorter a *simple possible world model* is defined to be a triple (W, B, I) such that (W, B) is a possible world model for belief in the sense of Section 1.1 and $I : \mathsf{At} \to \mathcal{P}W$ is an interpretation function. I am calling the possible world models defined here simple possible world models to distinguish them from the more complex models that are introduced later in the thesis.

As an example consider the model $(W, \{w\}, I)$, where I is the interpretation function defined above. According to this model the subject believes that it is raining and she is a speaker of English.

References to the literature

Carnap (1947) might be the first philosopher that uses something like the possible world framework to specify the meaning of sentences. Later, the possible world framework has become influential for the formal semantics of natural languages because of its use in Intensional Montague Grammar. I refer to Gamut (1991) and to Heim and Kratzer (1998) as introductions to the topic. In Montague semantics one usually employs logical languages of higher order that are far more complex than the propositional languages considered here. It is beyond the scope of this thesis to extend the account of interpretation to such higher-order languages.

My use of the interpretation function I in a possible world model (W, B, I) is different from the use it has in the literature on epistemic and doxastic logic mentioned in the previous section. In epistemic and doxastic logic the interpretation function, which is often called valuation function in that context, is thought of as specifying what basic facts hold at what worlds of the model. To do so one takes the meaning of the atomic sentences in At to be already understood. This is in some sense converse to the use of the interpretation function in this thesis. As in possible world semantics I assume that it is already understood in advance which basic facts hold at which worlds of the model. Given this knowledge one can use the interpretation function to capture the meaning of sentences in the language of the subject.

One can make two different purposes of interpretation functions explicit by considering models that contain two interpretation functions. Such a model is a quadruple $(W, B, I, I_\star)$. The first three components (W, B, I) are a simple possible world model as defined above. In particular $I : \mathsf{At} \to \mathcal{P}W$ is an interpretation function that represents the meaning of sentences in the language of the subject. The additional interpretation $I_\star : \mathsf{At}_\star \to \mathcal{P}W$ assigns meanings to sentences in a language that we, who are using the model to interpret the subject, already understand. It functions similarly to how the interpretation function is used in epistemic logic in that it describes the worlds in the model. It provides a formal specification of which basic facts hold at which worlds of the model.

Let us consider an example of such an extended model. Assume that $\mathsf{At} = \{p\}$ contains just one sentence p that stands for the expression "Pui pui." and that $\mathsf{At}_\star = \{q\}$ contains also just one sentence q that stands for the expression "It is raining." Consider now the model $M = (W, B, I, I_\star)$ where $W = \{w, v\}$, $B = \{v\}$, $I(p) = \{v\}$ and $I_\star(q) = \{w\}$. To understand what basic facts hold at the worlds in W we can use the interpretation $I_\star$. Because q is the sentence "It is raining." and we know that this sentence means in our language that it is raining we can conclude from $I_\star(q) = \{w\}$ that it is raining at the world w and that it is not raining at v. We can conclude that according to the model M the subject believes that it is not raining, and the sentence "Pui pui." means in her language that it is not raining. In this description of the model M I do not mention explicitly which basic facts obtain at which worlds of the model. Instead, I give the interpretation function $I_\star : \mathsf{At}_\star \to \mathcal{P}W$ and tell what the sentences in $\mathsf{At}_\star$ are supposed to mean.

In this thesis I am using models of the form (W, B, I) that do not contain an interpretation $I_\star$ for the language that we the modelers are using. Instead, I am assuming that when fixing the set of worlds W we already specify which basic facts hold at which worlds. This specification, which is given as part of the text describing the model, serves the same purpose as the interpretation function $I_\star$ would do in a model of the form $(W, B, I, I_\star)$.

1.3 The problem of radical interpretation

In this section I explain the problem that this thesis is concerned with and show that it is an instance of the problem of radical interpretation. I also discuss an analogy between the approach of this thesis and decision theory.

In Section 1.1 I give the intuitive explanation that a possible world is a doxastic alternative for some subject if and only if the world is compatible with everything the subject believes. To use this characterization for determining whether some possible world model represents the beliefs of some subject we need to assume that we already know in advance what the subject believes.

A similar point applies to the representation of meanings as propositions. In Section 1.2 I give the intuitive explanation that a possible world belongs to the proposition expressed by some sentence if and only if the sentence is true or false at the possible world. To use this characterization for determining whether some proposition captures the meaning of some sentence we need to know at which possible worlds the sentence is true. But it seems unlikely that we can come to know at which worlds some sentence is true without assuming that we already know what these sentences mean.

In this thesis I try to give an account of when a possible world model represents the beliefs of some subject and the meaning of her sentences that presupposes as little as possible prior knowledge about beliefs and meanings.

Let us consider two examples that give an idea of how one can find out about some of the subject's beliefs or about the meaning of some of her sentences. First suppose we want to find out whether the subject believes that it is raining. The obvious way to do so would be to just ask her "Is it raining?" If she says "yes" and is sincere then she believes that it is raining. In an even simpler case we might observe that the subject sincerely asserts the sentence "It is raining." In this case we can conclude that most likely she believes that it is raining. We can describe both versions of this example, the first one where the subject replies to our question and the second where the subject asserts a sentence herself, as cases where the subject accepts some sentence of which we already know that it means that it is raining. From this we can infer that she believes that it is raining. This however depends on our assumption that the accepted sentence means that it is raining. If the subject was using a language that is phonetically similar to English but assigns a different meaning to “It is raining.” then we could not conclude that she believes that it is raining. Also if the subject was to accept a sentence such as “Pui pui.”, of which we do not know what it means, then we are unable to conclude from this that she believes something as concrete as that it is raining.

Conversely, we can find out what the sentences in the language of the subject mean, assuming that we know what she believes. Let me explain this with another example. Suppose we want to determine whether the sentence “Pui pui.” means in the language of the subject that it is raining. Suppose that we know what the subject believes in various different situations. If we find that in every situation where the subject sincerely asserts the sentence “Pui pui.” she believes that it is raining and she never sincerely asserts “Pui pui.” if she does not believe that it is raining then this seems to make it quite plausible that “Pui pui.” means in her language that it is raining. In general, we might determine the proposition that is expressed by some sentence by observing which beliefs, thought of as propositions, are correlated with the acceptance of the sentence.

In both of these two examples one might say that we are interpreting the linguistic behavior of the subject. In the first example we are interpreting her reply to our question to deduce what the subject believes. In the second example we are interpreting her sincere assertion of some sentence in different situations to determine its meaning. This shows that a way of finding out what the subject believes and what the sentences in her language mean is to interpret her linguistic behavior. We could give an account of when a possible world model represents the beliefs of some subject and the meaning of sentences in her language by defining when a possible world model captures the result of interpreting the linguistic behavior of the subject.

The account of interpretation in this thesis should depend as little as possible on our prior judgments about beliefs and meanings. In the two examples above this is not the case, since we either presuppose that we know the meaning of the sentences that the subject is using to determine her beliefs, or we presuppose that we know what she believes in different situations to determine the meaning of one of her sentences. But ideally we would have an account of interpretation that does not assume any prior knowledge about the beliefs of the subject nor about the meaning of her sentences. Interpretation that does not presuppose any knowledge about beliefs and meanings is called radical interpretation. Hence, in this thesis I try to give an account of radical interpretation in which beliefs and meanings are represented with possible world models.

In radical interpretation we try to determine the beliefs of some subject and the meaning of sentences in her language from the linguistic behavior of the subject. The crucial aspect of the subject’s linguistic behavior is, in both examples above, what sentences she accepts. Intuitively, one might say that the subject accepts a sentence in some situation if and only if she is disposed to sincerely assert that sentence or she would reply affirmatively if we presented the sentence as a question. I do not further analyze the notion of acceptance on its own, but take it as a primitive for the account of interpretation given in this thesis.

The notion of acceptance of sentences fits well with the representation of belief and meaning in possible world models. The bridge between the acceptance of sentences and possible world models is the following *acceptance principle*:

> The subject accepts a sentence if and only if she believes the proposition expressed by the sentence.

This acceptance principle plays a central role in the account of interpretation developed in the first part of this thesis. The second part of the thesis explores the consequences of two distinct adaptations of the simple acceptance principle given here to more complex notions of meaning.

The notion of acceptance alone is not enough to yield a satisfactory account of interpretation. Even the most radical interpretation considered in this thesis assumes that we know some of the subject's beliefs prior to interpretation. It is assumed that we know all the beliefs that the subject obtains from perceiving her situation in the world. I call them the perceptual beliefs of the subject in some situation. As an example we might observe the subject walking through the rain and getting wet. If we make the assumption that the subject is a sensible human being we can conclude that the subject experiences that it is raining and hence believes that it is raining. To determine that in this situation the subject believes that it is raining we do not need to interpret her linguistic behavior.

The approach to the problem of radical interpretation taken in this thesis is methodologically similar to the approach in decision theory. Decision theory uses the subject's choice behavior to determine whether some probability distribution and utility function represent the beliefs and desires of the subject. The subject's choice behavior is thought of as the choices that the subject is making between the different acts that are available to her in some situation.

One can use the subject's choices between different acts to determine some of her beliefs. For instance we might find that before leaving the house the subject is putting on her rain jacket instead of her pullover and conclude from this that she believes that it is going to rain. This presupposes however that we know that the subject dislikes getting wet in the rain and would usually find it more convenient to wear her pullover than her rain jacket. If the subject would not mind getting wet, but very much liked the color of her rain jacket, then we could not conclude from the fact that she is putting on her rain jacket that she believes that it is going to rain. To give an account of the subject's beliefs in terms of her choice behavior decision theory needs a formal representation of the subject's desires, which is provided by the utility function. The role of desire in decision theory is similar to the role of meaning on the account of interpretation sketched above, in that together with the subject's beliefs they are used to interpret the subject's behavior.

The acceptance principle introduced above plays a similar role as the principle of expected utility maximization in decision theory. Expected utility maximization requires that the subject chooses one act over another if and only if the former has a higher expected utility than the latter. This principle connects the subject's choice behavior to the probability and utility functions that represent her beliefs and desires. Similarly, the acceptance principle connects the subject's linguistic behavior to the belief set and the interpretation function that represent her beliefs and meanings in her language.

Also the assumption that we know at least the perceptual beliefs of the subject has a counterpart in decision theory. There it is assumed that we know what consequences the subject believes her acts to have in different states of the world. For this we need to know some of the basic beliefs that the subject has about her environment. If the subject is putting on her rain jacket when leaving the house this only relates to her credence for rain if we can make the assumption that she knows that it is the rain jacket that she is putting on. If the subject was to confuse her rain jacket with her pullover then her putting on her rain jacket would not show that she assigns a high probability to rain.

A difference between the setting of the thesis and decision theory is that in decision theory the beliefs of the subject are represented by subjective probabilities. This makes it possible to weigh the strength of beliefs against the comparative strength of desires, which is needed to apply the principle of expected utility maximization. The only thing that we can conclude when observing the subject putting on her rain jacket is that her belief that it is going to rain is at least as strong as her dislike for wearing a rain jacket relative to her dislike for getting wet when wearing a pullover. If the subject does not particularly dislike wearing a rain jacket but very much dislikes getting wet in a pullover then she might put on her rain jacket, even though she thinks it is less probable that it rains than that it does not. In this situation we could even imagine that the subject accepts the sentence "It is not going to rain." with its usual meaning in English. Following the acceptance principle we would conclude that she has the belief that it is not going to rain. However, for the principle of expected utility maximization the remaining probability of the belief that it is going to rain is high enough to make her put on a rain jacket because she wants to avoid the unlikely but very unpleasant event of getting into rain with just a pullover.

References to the literature

The interest in the problem of radical interpretation originates from its extensive discussion by Quine (1960), who calls it the problem of radical translation. Following Quine, Davidson (1973; 1974; 1975) has worked extensively on the problem of radical interpretation. The approach of this thesis is strongly influenced by this work. But I also refer to the work on radical interpretation by Lewis (1974) and McCarthy (2002).

Throughout this thesis I refer to the literature mentioned in the previous paragraph to compare it with the approach of the thesis. There are however three points I want to mention already here.

First, in basing my account on the notion of the subject accepting a sentence I am following Davidson (1973). The notions of accepting a sentence is intended to be similar to Davidson's notion of holding a sentence true. Lewis (1974) assumes that interpretation is based on our prior knowledge of all facts about the subject as a physical system. This is wider than the notion of acceptance in that it also comprises the subject's brain state or all of the behavior that is not linguistic. It is also more elementary than acceptance because to see whether the subject accepts some sentence we need to assume for instance that the subject is sincere or that she understands that she is making an assertion. McCarthy (2002) also seems to use a wider base for interpretation that includes all of the subject's behavior. On both Lewis' and McCarthy's account interpretation is supposed to yield knowledge about all of the subject's attitudes, which also include her desires. This is far more ambitious than the account in this thesis which only includes beliefs. One advantage of using the notion of acceptance as basic is that it abstracts away from the subject's motivation for her linguistic behavior.

Second, a crucial part of existing accounts of radical interpretation is to determine the compositional structure of the subject's language and the meaning of subsentential expressions. The resulting problem of the inscrutability of reference has inspired a lot of the literature on radical interpretation (see for instance Quine 1968; Davidson 1979). McCarthy's (2002) account is mainly focused on this problem. As mentioned in Section 1.2 I am only considering propositional languages which are not compositional on the subsentential level. Hence, I am not engaging with the problem of inscrutability of reference.

Third, none of the accounts of radical interpretation mentioned above uses the possible world framework to model belief and meaning. A reason for this might be that Davidson, who wrote the most about the problem, inherited Quine's suspicion against propositions and possible worlds. I do not share this suspicion and my goal in this thesis is to use the setting of radical interpretation to make sense of the possible world framework for belief and meaning.

Stalnaker (1984) has a similar ambition as this thesis in that he aims to vindicate the possible world framework for propositional attitudes and meaning. Stalnaker (1984, p. 36) dismisses a theory of interpretation inspired by Davidson as part of "the linguistic picture" of intentionality that he rejects. Stalnaker's own account, "the pragmatic picture", models propositional attitudes and meanings using possible worlds. He does not, however, present a theory that is systematic enough to be formalized. Stalnaker (1984, pp. 17–18) suggests that the ascription of beliefs should be related to our ability to explain the actions of the subject and is sensitive to what causes the beliefs under optimal condition, where these optimal conditions might depend on social and linguistic factors (p. 67).

In later work Stalnaker (for instance 1990, p. 144) explicitly endorses an information-theoretic account, along the lines of Dretske (1981), to explain how the contents of beliefs are determined. On this account beliefs are thought of as mental representations that play a causal role in the subject considered as a biological organism. The content of such mental representations is determined by their evolutionary function. I do not know whether these ideas can be turned into a systematic theory that explains when a possible world model represents the beliefs of the subject and the meaning of her sentences. In this thesis, I however employ a different approach that does not consider the causal role of mental representations, but rather takes beliefs to be a theoretical notion that helps to explain linguistic behavior.

As an introduction to decision theory I refer to Jeffrey (1983) who presents a version of decision theory widely used in philosophy. Savage (1972) provides the original treatment of decision theory.

The connection between the theory of interpretation and decision theory has been noticed in the literature. The relation between the two however is not entirely clear. Lewis (1974, p. 337) thinks that the ascription of beliefs and desires that is the outcome of interpreting the subject needs to be compatible with a decision-theoretic explanation of the subject's choice behavior. For Lewis decision theory is a part of the theory of interpretation that constrains possible ascriptions of beliefs and desires to the subject. Davidson (1980) suggests to combine decision theory with a theory of interpretation for the language of the subject by using the notion of the subject preferring the truth of one sentence to the truth of another sentence as the basic notion of behavior.

In this thesis I treat the theory of interpretation as methodologically analogous to but distinct from decision theory. The two theories concern different abstract notions. Decision theory concerns probabilistic beliefs and desires. The account of interpretation given in this thesis concerns qualitative beliefs and meanings.

1.4 Outline of a solution

In this section I explain the general strategy that I use to solve the problem of radical interpretation. I also introduce three requirements on an account of interpretation that help evaluating and comparing different accounts.

In this thesis I consider different accounts of interpretation that vary in the details of the modeling. The accounts have in common that they all consist of five steps.

In the first step, I define the class of possible world models that are used to represent beliefs and meanings. In the simplest case these are the simple possible world models from Section 1.2. But I also consider more sophisticated models that use more refined structures to represent belief and meaning.

The second step is a definition of some notion of linguistic behavior. In the most basic case a linguistic behavior is a set of sentences in the language of the subject, which we think of as containing all the sentences that the subject accepts in some situation.

In the third step, I give a definition of the linguistic behavior that is generated by some possible world model. If a model generates the linguistic behavior of some subject then we can think of the model as representing the beliefs of the subject and the meaning of sentences in language. In this case I also say that the model interprets the linguistic behavior of the subject. In all accounts of this thesis the definition of the behavior generated by a model is going to be an adaption of the acceptance principle from Section 1.3 to the notion of possible world model and linguistic behavior used by the account.

In the fourth step, I specify what additional knowledge about the subject, besides knowing her linguistic behavior, we assume to have in advance to interpretation. We might for instance assume that we know the meaning of some of the expressions that the subject is using or that we know some of her beliefs.

Carrying out these first four steps yields a concrete formal account for interpretation. A linguistic behavior is interpretable according to the account if it is the linguistic behavior that is generated by some model that satisfies the additional assumptions made in the fourth step. There is one further fifth step that characterizes the class of interpretable behaviors.

In this fifth step, I prove a representation result that gives necessary and sufficient conditions on a linguistic behavior to be interpretable according to the account that is set up in the first four steps. These representation results allow us to evaluate and compare different accounts of interpretation by considering the conditions they place on interpretable behaviors.

I use three requirements to evaluate a framework of interpretation which is set up according to the five steps described above. The requirements capture that the formal account of interpretation should allow us to unambiguously and radically interpret all linguistic behaviors.

The first requirement is the following *variety requirement* which ensures that we can interpret all linguistic behaviors:

> Every linguistic behavior that some subject might plausibly show should be interpretable.

We do not want to exclude some plausible linguistic behaviors from the account just because our models are too restrictive. To evaluate an account of interpretation on this requirement one can use the characterization of interpretability in the representation results from the fifth step above. The account fulfills the requirement if one can argue that every plausible linguistic behavior satisfies the conditions for interpretability that are given by the representation result. To show that the account does not satisfy the requirement one needs to find a plausible behavior that does not satisfy the conditions of the representation result.

One can also consider the converse of the variety requirement, which says that every behavior that is interpretable according to the formal account should also be a behavior that some subject might plausibly show. Although this converse version of the variety requirement is interesting it does not play an important role in this thesis.

The second *determinacy requirement* ensures that interpretation leads to an unambiguous result:

> A linguistic behavior should be interpretable by at most one model.

We want that according to all the models that interpret one behavior the subject has the same beliefs and the meanings of her sentences are the same. The statement of the determinacy requirement does not specify what is meant by one model because it leaves it open when two models count as the same. In this thesis I do not apply the determinacy requirement very strictly and hence I do not give a formal definition of when two possible world models count as the same. I only invoke the determinacy requirement against accounts of interpretation that lead to an obvious indeterminacy of almost all beliefs or meanings.

The third requirement ensures that interpretation does not rely on prior knowledge about belief and meaning. I call it the *little-input requirement*:

> The prior knowledge about the subject that is assumed by the account should be available to a radical interpreter.

This requirement says that the account of interpretation should do as much work as possible. Ideally, we infer the subject's beliefs and the meaning of her sentences from nothing but her linguistic behavior.

Developing the account of interpretation is a matter of balancing these three requirements. For instance one is often tempted to enlarge the class of interpretable behaviors by refining the structures representing belief and meaning in possible world models. However, this usually increases the indeterminacy, unless one is assuming additional prior knowledge about the subject.

The aim of this thesis is to find a formal account of interpretation that performs reasonably well on the three requirements. This task is difficult enough even with a lot of idealizing assumptions and using only the most simple formal models for belief and meaning. I am not attempting to do more than this. Let me explicitly mention three things that I am not doing in this thesis.

First, I am not aiming for an account that does justice to our common-sense theory of belief and meaning. As a consequence I do not take the intuitive appeal of some formal model to be a reason for or against using it. I only judge a formal model on whether it leads to an account of interpretation that performs well on the three requirements stated above.

Second, I am not giving a formal semantics of belief ascriptions in natural language. One might hope to gain insights into the interplay between the notions of belief and meaning by investigating belief ascriptions. But this is not the route taken in this thesis. If the linguistic behavior of the subject includes belief ascriptions to other subjects then the formal semantics of belief ascriptions might become relevant for the theory of interpretation. This is however far beyond the scope of this thesis.

Third, I am not trying to provide a formal model of what is happening inside human brains when we have certain beliefs or utter certain sentences. This thesis rather provides a conceptual investigation of the possible world framework for belief and meaning.

Developing an account interpretation of the kind outlined above provides a precise characterization of the assumptions that are made by some kind of possible world models for belief and meaning. The representation result yields conditions on linguistic behaviors that capture the class of behaviors that the kind of models under consideration can account for. If one then finds that these constraints are unsatisfactory, for instance because actual linguistic behavior does not satisfy them, then one can adapt the formal models to obtain more plausible conditions.

References to the literature

My approach to solving the problem of radical interpretation is methodologically analogous to the approach taken by many presentations of decision theory. Most presentations contain all of the five steps outlined above. The first step corresponds to the choice of the mathematical structure that represents the subject's beliefs and utilities. The second step introduces a notion of the subject's choice behavior which is usually a preference relation or selection function over acts. In the third step, a decision rule is given, such as for instance the principle of expected utility maximization or the maximin rule. The fourth step is often not made explicit. but there are also many version of decision theory where for instance we assume that the subject knows the objective probabilities or where we assume to know the subject's utilities. The fifth step corresponds to the representation theorems in decision theory that characterize those choice behaviors that arise from a formal model of beliefs and utility using some choice rule.

The three requirements that I use to evaluate an account of interpretation can be found in one form or other at various places in the literature.

Something like the variety requirement is usually not mentioned in the literature on the problem of radical interpretation. The reason might be that this literature does not try to represent belief and meaning with some class of formal models that might turn out to be too restrictive. In decision theory an analogue of the variety requirement is employed when a framework for decision making is evaluated on the plausibility of the axioms required by the representation theorem. Alternative models for decision making are developed because experiments have shown that the choices of actual agents do not fulfill the axioms of classical decision theory. In formal semantics something similar to the variety requirement is in play when an existing account is extended in order to account for a richer class of sentences or inferences. In this context it seems that also the converse of the variety requirement is used when a semantic theory is criticized for taking a certain sentence or inference to be semantically admissible that intuitively is not.

The determinacy requirement plays a prominent role in all of the work on radical interpretation. Quine (1960) questions the possibility of a scientific theory of meaning because some indeterminacy is inevitability. Lewis (1974) suggests that all indeterminacy might be eliminated and that an account of interpretation is not complete until it does so. Davidson (see for instance 1974, pp. 153–154) accepts that some indeterminacy of interpretation can not be avoided and considers this to be a feature and not a defect of the theory of interpretation. In this thesis I have similar attitude and only worry about the problem of indeterminacy if it is so pervasive that interpretation completely fails at constraining the subject's beliefs and the meaning of the sentences that she is using.

Existing accounts of radical interpretation have an analogue of the little-input requirement in that they constrain what evidence about the subject is assumed to be available to an interpreter. Quine (1960) assumes knowledge about the subject's sensory stimulation and her verbal behavior. Lewis (1974) takes all facts about the subject as a physical system as given for interpretation. Davidson seems to base his theory of interpretation on all the information about the subject that is publicly observable. In this thesis I am focusing, like Davidson, on the public availability of the information given to the interpreter rather than requiring that it can be described in a purely physical language. I do not apply the little-input requirement very strictly and also consider accounts of interpretation that presuppose that we know in advance some of the subject's beliefs or the meaning of some expressions in her language. This is helpful since it allows us to start from a simple account and then iteratively improve it to assume less prior knowledge about beliefs and meanings.

Davidson (1973) emphasizes the further requirement that the interpreter has only a finite amount of evidence available for interpretation and that the theory of meaning for the subject's language should be expressible in a finite description. Davidson suggests that these constraints entail, together with the additional assumption that the subject's language contains infinitely many sentences, that the theory of meaning needs to exploit, and hence account for, the compositional structure of the subject's language. I do not use this requirement in this thesis because I restrict myself to modeling the sentences in the language of the subject with propositional formulas, whose compositional structure is too poor to generate infinitely many non-equivalent sentences from a finite set of primitives.

1.5 Overview of this thesis

In the following I give an overview of the remaining chapters.

In Chapter 2 I develop a first account of interpretation that serves as the basis for the accounts in later chapters. The account is simple but it makes the strong assumption that we know all of the subject's beliefs prior to interpretation.

In Chapter 3 I develop an account of interpretation on which it is only assumed that we know the perceptual beliefs that the subject has about the world and it is left to the account of interpretation to deduce her non-perceptual beliefs. The resulting setting uses a more complex model of belief than the one from Chapter 2. Compared to all other accounts discussed in this thesis it performs best on the three requirements from Section 1.4. Readers that are mostly interested in formal semantics and the modeling of meaning can safely skip Chapter 3, because later parts of the thesis only use the setting of Chapter 2.

Chapters 4, 5 and 6 are concerned with modeling of meaning and the role that meaning plays in interpretation.

Chapter 4 introduces a distinction between two different versions of the acceptance principle that profoundly influences the structure of the more complex representations of meaning used in Chapters 5 and 6.

In Chapter 5 I compare the possibilities of accounting for vague expressions in the language of the subject as resulting either from an indeterminacy in the semantic facts or from an uncertainty in the beliefs of the subject about the semantic facts. This discussion is relevant for the overall theory of interpretation because it provides an explanation for a technical construction, which involves the splitting of possible words, that is already introduced in Section 2.4 but seems somewhat unnatural at that point.

Chapter 6 treats interpretation in the case where we have a hypothesis about what parts of the language of the subject function as a necessity modality. I discuss two different formal models that correspond to different interpretation of two-dimensional semantics.

Readers that are interested in doxastic logic and do not care much about meaning can skip the latter three chapters unless they are concerned about the splitting of worlds in Section 2.4. In the latter case minimally the material from Sections 4.1, 4.3 and 5.1 is needed to address this concern.

Chapter 7 provides the mathematical background for the thesis. It contains the proofs for the representation results mentioned in the earlier chapters and some additional definitions and examples that are too technical to be discussed as part of the main text.

The results of this thesis are original work by the author and have not been published before.

Chapter 2

The framework

In this chapter I develop a first account of interpretation. I start in Section 2.1 with a very simple account that introduces the relevant concepts but performs badly on the determinacy requirement. I then in Section 2.2 and 2.3 discuss two ways to improve the account by either assuming that we know the subject's language or that we know her beliefs. In this thesis I focus on the latter approach. The setting from Section 2.3 however still has difficulties with the variety and determinacy requirements. In Sections 2.4 and 2.5 I show how these problems can be solved. This leads to the account of interpretation from Section 2.6 which is the bases for the discussion in later chapters.

2.1 A first attempt

In this section I give a first simple account of interpretation. The account has serous difficulties with the determinacy requirement but it is the basis for the improvements in later sections of this chapter.

Let us suppose that we are in some situation where we want to find out what some subject believes and what the sentences in her language mean. We assume that the only thing that we know about the subject is the set of sentences which she accepts in this situation. This is captured by the following definition of a linguistic behavior:

2.1.1. Definition. A *linguistic behavior* is a set $A \subseteq \mathcal{V}$ of sentences.

I often just write *behavior* when I mean a linguistic behavior. We think of a linguistic behavior A as the set of all the sentences which the subject accepts in the situation where we are interpreting her. Hence for every $\varphi \in \mathcal{V}$ it should hold that $\varphi \in A$ if and only if the subject accepts φ.

As an example imagine a situation where the subject does not assert the sentence “It is raining.” and maybe she expresses doubts if we present her the sentence “It is raining.” She does however sincerely assert the sentence “There are raindrops on the window." To formalize this let p be the sentence "It is raining.", q the sentence "There are raindrops on the window." and assume that set $\{p, q\} = \mathsf{At} = \mathcal{V}$ is the set of all sentences in the vocabulary of the subject. We can model the example as a situation where the subject does not accept p but accepts q. Her linguistic behavior in this situation is the set $A = \{q\}$.

We now define what linguistic behavior the subject shows if a certain simple possible world model correctly describes her beliefs and the meaning of her sentences. According to the acceptance principle the subject accepts a sentence if and only if she believes the proposition that the sentence expresses. In a simple possible world model (W, B, I) the subject's beliefs are represented by the belief set $B \subseteq W$ and the meaning of her sentences are given by the interpretation function $I : \mathcal{V} \to \mathcal{P}W$. Relative to this representation the acceptance principle requires that the subject accepts a sentence $\varphi \in \mathcal{V}$ if and only if $B \subseteq I(\varphi)$. This yields the following definition:

2.1.2. Definition. The *linguistic behavior* $A^M \subseteq \mathcal{V}$ *generated* by the simple possible world model $M = (W, B, I)$ is defined by:

$$A^M = \{\varphi \in \mathcal{V} \mid B \subseteq I(\varphi)\}.$$

As an example consider the model $M = (W, B, I)$ where $W = \{w, v, u\}$, $B = \{w, v\}$ and $I(p) = \{w\}, I(q) = \{w, v\}$ and assume that $\mathcal{V} = \mathsf{At} = \{p, q\}$. We think of w as a world where it is raining and there are raindrops on the window, v as a world where it is not raining but there are still raindrops on the window and u as a world where it is neither raining nor are there raindrops on the window. According to the model M the subject believes that there are raindrops on the window and is uncertain about whether it is raining. The meanings of the sentences in her language are just the meanings that they have in English, if we take p to be "It is raining." and q "There are raindrops on the window." The behavior generated by this model is $A^M = \{q\}$.

The behavior of the subject is interpretable by the formal account of interpretation of this section if it is the behavior generated by some model. This leads to the following definition:

2.1.3. Definition. A simple possible world model M *interprets* a linguistic behavior A if $A = A^M$.

A linguistic behavior $A \subseteq \mathcal{V}$ is *interpretable* if there exists some simple possible world model M that interprets A.

The behavior A from the example at the beginning of this section is interpretable because it is interpreted by the model in the example from the paragraph above.

I continue by characterizing the class of interpretable behaviors. First consider the case where $\mathcal{V} = \mathsf{At}$, that is, we take all the sentences in the language of the subject to be atomic and hence are not making any assumptions about which linguistic constructions in her language correspond to the propositional connectives. It is a simple observation that in this case any behavior $A \subseteq \mathsf{At}$ is interpretable. A formal proof of this fact is provided by Proposition 7.1.1 in Chapter 7. The requirement of being interpretable with simple possible world models does not place any constraints on linguistic behaviors.

The notion of interpretability becomes more interesting if we already have a hypothesis about the propositional connectives in the language of the subject. In this case we take the subject to be accepting formulas in propositional logic, and hence we set $\mathcal{V} = \mathcal{B}$. Proposition 7.1.2 shows that a linguistic behavior $A \subseteq \mathcal{B}$ is interpretable if and only if it is a propositional theory. As explained in Section 1.2 this means that the set of sentences that the subject accepts needs to be closed under logical consequence.

Let us now see how the simple account outlined here fares with respect to the requirements on a theory of interpretation from Section 1.4.

We first consider the little-input requirement. We do not suppose any knowledge about the subject's beliefs or the meaning of her sentences. The account only requires that we know the set of sentences that the subject accepts. I take this to be a rather weak assumption that is reasonable for a theory of interpretation. However, let me discuss two reasons for which one might think that it is already too strong to require that for every sentence in the language of the subject we know whether the subject accepts this sentence.

First, an actual subject might not utter a sentence even though she accepts it, because she has no reason to do so. Or, she might utter sentences which she does not accept because she wants to deceive someone. In both cases the interpreter would have difficulties determining which sentences she accepts. Accounting for such cases seems to require that we model the intentions that cause the linguistic behavior of the subject. Because I do not do this in this thesis I just assume that we know which sentences the subject accepts.

Second, one might object that an actual interpreter can fail to observe which sentences the subject accepts in some situation, even under the assumption that she utters precisely those sentences that she accepts. Maybe the interpreter just does not pay enough attention to the subject. I do not think that this problem should be accounted for by the theory of interpretation. The theory of interpretation only tells us how from a given linguistic behavior we can derive a formal model of the subject's beliefs and the meaning of sentences in her language. If an interpreter fails to gather enough information about the subject's linguistic behavior then this is not a problem for the theory of interpretation. It simply leads to an uncertainty in the interpreter's knowledge about the subject which could be accounted for in a higher-order account of interpretation of interpreters.

In the case of propositional sentences there is one further reason why the account given here might not perform well on the little-input requirement. The problem is that in this case the account describes how to interpret a subject given a hypothesis about what expressions in the subject's language play the role of the classical propositional connectives. It would be desirable if the account would not make this assumption but would detect the propositional connectives in the subject language as part of interpretation. I have no idea how such an account could be made to work and hence, throughout this thesis, I do assume that already prior to interpretation we have a hypothesis about the propositional connectives in the language of the subject.

With respect to the variety requirement I take the account of this section to work reasonably well. If we only consider the acceptance of atomic sentences then our representation result shows that every linguistic behavior is interpretable. This seems right, since for any subset of the set of atomic sentences we can imagine the meanings of the sentences in the language of the subject to be such that exactly the sentences in the subset mean something that the subject believes.

For propositional sentences we have that to be interpretable a linguistic behavior needs to be a theory of classical propositional logic. I take this to be reasonable requirement and do not discuss any possibilities for weakening it in this thesis. Let me however mention two kind of examples in which the subject's linguistic behavior is not a theory of classical logic.

First, we can imagine that the subject systematically violates the constraint that the set of sentences she accepts is a theory. Such an example would show that our initial hypothesis about which constructions in the language of the subject correspond to the classical propositional connectives is mistaken. It would not provide a reason to think that the account of interpretation given here does not satisfy the variety requirement.

It could be that the subject uses a language in which no constructions play the role of the classical propositional connectives. In this case no hypothesis about what the propositional connectives in the subject's language are renders her linguistic behavior interpretable. The account for interpretation considered here is of no use for such a subject. There might then still be expressions in the subject's language which function as the propositional connectives for some alternative logic other than classical logic. The subject might for instance be an intuitionist or she might accept contradictions. In this thesis I restrict myself to the setting of classical logic but it would be interesting to see whether an account of interpretation similar to the one given here could be developed for other logics.

Second, the subject might violate the constraint that the set of sentences which she accepts is a theory in certain exceptional cases where she fails to notice that a certain sentence follows in classical propositional logic from the sentences that she already accepts. Such divergences from classical logic would need to be exceptional because if they were not then we would rather take them as evidence against our hypothesis about which constructions in the subject's language correspond to the classical connectives. To account for such failures to draw all logical inferences one would need a formal model of the subject's restricted computational capacities. I do not consider any such models in this thesis.

The most severe problem for the account of interpretation given in this section is that it does not fulfill the determinacy requirement. The account does not specify what basic facts obtain at the worlds in the domain of a model that interprets the behavior of the subject. As a consequence we can not make sense of the worlds in the belief set to determine what the subject believes and we can not deduce the meaning of sentences in the language of the subject from knowing the sets of worlds where they are true.

Consider an example. Let us assume that the subject accepts the sentence p and all its consequences in propositional logic such as $p \vee q$, $\neg\neg p$ or $\neg p \to q$. We can capture this linguistic behavior with the set $A = \mathsf{cl}(\{p\}) \subseteq \mathcal{B}$. Since A is a theory we know that there is some model $M = (W, B, I)$ such that A is the behavior A^M generated by M. This model should tell us what the subject believes and what the sentences in her language mean. Let us for instance check whether the subject believes that it is raining. According to a possible world model the subject believes a proposition if it is true at all worlds in the belief set of the model. But we can not tell whether it is raining at the worlds in B because we do not know which basic facts obtain at the worlds in W. The model M satisfies the definition for interpreting the behavior A independently of how we think of the worlds in its domain. For the same reason we can also not use M to determine the meaning of sentences in the subject's language. For instance we might wonder what the meaning of p is. From possible world semantics we know that it is the set $I(p) \subseteq W$. But this does not help since we have no idea what the worlds in $I(p)$ look like. At these worlds any basic facts might hold and consequently p might mean anything.

For a different perspective on this problem consider the model $M = (W, B, I)$ that is defined in the proof of Proposition 7.1.2 which shows that every behavior that is a propositional theory is interpretable. The worlds in the domain W of M are defined to be all the complete theories in the language of he subject. The interpretation function $I : \mathsf{At} \to \mathcal{P}W$ sends an atomic sentence p to the set of complete theories that contain p. This does not tell us anything about the meaning of p. The possible worlds which are supposed to fix the meaning of sentences in the subject's language are themselves linguistic constructions from sentences in this language. As a consequence we are also not able to tell what the subject believes. The belief set B is defined to contain all the complete propositional theories which extend the theory A of sentences that the subject accepts. Because we do not know what these sentences mean we can not tell what facts obtain at the worlds that the subject considers possible.

Let me explain in what sense the problem explained in the previous paragraphs can be seen as a case where an account of interpretation does not satisfy the determinacy requirement. The problem is that the account does not determine what basic facts obtain at the possible worlds in the model that interprets a linguistic behavior. It is not obvious why this should be taken to be a violation of the requirement that every behavior should be interpreted by a single model. If we take models to be mathematical structure of the form (W, B, I) then the problem is not that there are many distinct such triples that interpret the behavior of the subject. The problem is that we do not know how to think about the worlds in W. But we can also take a model to be a triple (W, B, I) together with an informal description of which worlds satisfies which basic facts. Then there are many different models that interpret some behavior because the account of interpretation does not fix such an informal description.

One can make this indeterminacy mathematically explicit by considering possible world models that are quadruples $(W, B, I, I_\star)$ as described on page 9 at the end of Section 1.2. The reduct (W, B, I) of such quadruples is a simple possible world model and $I_\star : \mathsf{At}_\star \to \mathcal{P}W$ is an additional interpretation function that maps atomic sentences in the language of the modeler to sets of worlds in order to determine which basic facts obtain at which worlds of the model. The problem of the account of interpretation discussed here is that nothing determines the interpretation $I_\star$. If a simple possible world model $M = (W, B, I)$ interprets some linguistic behavior then any extended model $(W, B, I, I_\star)$ based on M also interprets the linguistic behavior, no matter what $I_\star$ is.

In the remaining sections of this chapter, I discuss different improvements of the account of interpretation given in this section that avoid the indeterminacy explained. To fix the indeterminacy we need to make sure that during the process of interpretation we come to know which basic facts hold at the worlds of a model that interprets the behavior of some subject. One way of doing this is to already determine a set of worlds W prior to interpretation and to require that an interpreting model has W as its domain. Because we could then already specify in advance which basic facts hold at which worlds in W we would have no problems to make sense of the interpreting model. The task of interpretation is then to relate the linguistic behavior of the subject to the given domain W.

In the following two sections, I discuss two possibilities for relating the linguistic behavior of the subject to a domain W that is given in advance to interpretation. In both approaches I assume that we posses additional knowledge about the subject. The first approach, explained in Section 2.2, presupposes that we know already in advance what propositions, as subsets of the given domain W, the sentences in the vocabulary of the subject express. The other approach, introduced in Section 2.3 and improved in Sections 2.4, 2.5 and 2.6 presupposes that we know the belief set, as a subset of W, in advance. In both approaches we need very strong assumptions about the interpreters prior knowledge to solve the problem of indeterminacy described here. In Chapter 3 I explain how to improve the second approach such that it only assumes that we have knowledge of the subject's perceptual beliefs instead of assuming that we know all of her beliefs.

References to the literature

The difficulties that the account of this chapter faces with respect to the three requirements have all already been considered in the literature.

The problem of identifying the connectives of classical propositional logic in the language of the subject is discussed by Quine (1960, sec. 13). Quine states criteria that the propositional connectives in the language of the subject have to satisfy. For instance the conjunction needs to have the property that whenever the subject accepts a conjunction of two sentence then she also accepts the two sentences individually. To identify the propositional connective in the subject's language we have to find expressions which satisfy these criteria.

Quine does not tell us how we might come up with a hypothesis about which expressions of the subject's language are the propositional connectives. He just provides us with criteria to verify a given hypothesis. In this sense Quine's procedure for translating the propositional connectives is similar to the account given in this section. Quine's criteria on the translation of propositional connectives are formalized in this section by the requirement that the set of sentences that the subject accepts in one situation is a theory of classical propositional logic.

The difficulty for the variety requirement that the set of sentences that the subject accepts is supposed to be closed under logical consequence is closely related to the problem of logical omniscience in epistemic and doxastic logic. Logical omniscience is the property that the subject's beliefs are closed under logical consequence or the implication of propositions. For an overview of different approaches for dealing with logical omniscience in epistemic and doxastic logic I refer to (Fagin et al. 2003, ch. 9) and (Halpern and Pucella 2011).

In the literature on logical omniscience it is not always clear whether the problem concerns sentences or propositions. Most authors working on doxastic logic take the problem to be one about beliefs in sentences and hence are worried about the closure of these beliefs under logical consequence. This does not quite match the setting of this thesis because I carefully distinguish between sentences and propositions and take belief to be an attitude towards propositions and not sentences. The problem for the account of this section is rather that acceptance, which is an attitude towards sentences, is not closed under logical consequence.

A different version of the problem of logical omniscience concerns the closure of beliefs under the implication of propositions. This version of the problem is the one addressed for instance by Stalnaker in (1984, ch. 5) and in (1999). The variety requirement gives us no reason to worry about the closure of beliefs under implication because it only concerns the subject's linguistic behavior which is defined in terms of the acceptance of sentences. The notion of a belief in a proposition is internal to the theory and hence not affected by the requirement. It might however turn out that the best solution to the problem that acceptance of sentences is closed under logical consequence also requires us to give up closure of belief under implication of propositions.

The difficulty with the determinacy requirement discussed at the end of this section can be seen as a variation on Putnam's paradox (see Putnam 1981, ch. 2), especially in the rendering of the paradox by Lewis (1984). The main difference between Putnam's paradox and the problem discussed above is that Putnam's paradox concerns the reference of words to objects and properties whereas the problem discussed above is about the propositions expressed by sentences.

2.2 Fixing meanings

In this section I discuss an account of interpretation on which it is assumed that we already know in advance what the sentences in the language of the subject mean. This assumption solves the problem from the previous section because we are interpreting the subject with a model that is based on a domain which we have fixed in advance when specifying the meaning of sentences in the subject's language. In the end of this section I suggest that the assumption that we have prior knowledge of the meaning of the subject's sentences might be too strong because it conflicts either with the little-input or with the variety requirement.

In this section I again take a linguistic behavior to be a set of sentences $A \subseteq \mathcal{V}$, which is thought of as the set of sentences that the subject accepts in some situation. Also the notion of a behavior generated by a model $M = (W, B, I)$ stays the same as in the previous section, that is, A^M is the set of all propositional sentences φ such that $B \subseteq I(\varphi)$.

What changes in this section is the notion of interpretability. We assume to have prior knowledge about the meaning of the sentences in the language of the subject. As explained in Section 1.2 we can represent the meaning of sentences with an interpretation function $I : \mathsf{At} \to \mathcal{P}W$, which maps an atomic sentence to the set of worlds where the sentence is true. The assumption that we have prior knowledge of the meaning of sentences in the subject's language corresponds to the assumption that we know the interpretation function of any model interpreting the subject. This motivates the following definition of interpretability:

2.2.1. Definition. A belief set $B \subseteq W$ *interprets a linguistic behavior* $A \subseteq \mathcal{V}$ *with an interpretation function* $I : \mathsf{At} \to \mathcal{P}W$ if $A = A^M$ for the simple possible world model $M = (W, B, I)$.

A behavior $A \subseteq \mathcal{V}$ is *interpretable with an interpretation function* $I : \mathsf{At} \to \mathcal{P}W$ if there is some belief set $B \subseteq W$ that interprets A with I.

Let us consider an example in which $\mathcal{V} = \mathsf{At} = \{p, q\}$. We start from an interpretation function that we assume to give the correct meaning of the sentences in the language of the subject. To define the interpretation we first need to fix the set of worlds that we are using as the domain. As in the examples of the previous section we use the set $W = \{w, v, u\}$, which contains three worlds: w where it is raining and there are raindrops on the window, v where it is not raining and there are raindrops on the window and u where it is neither raining nor are there raindrops on the window. Now let us assume that we know that in the subject's language p means that it is raining and q means that there are raindrops on the window. Hence the interpretation function $I : \mathsf{At} \to \mathcal{P}W$ for the subject's language is given by $I(p) = \{w\}$ and $I(q) = \{w, v\}$.

Now suppose that the subject shows the linguistic behavior $A = \{q\}$. A model that interprets this behavior with I is $M = (W, B, I)$, where $B = \{w, v\}$. According to this model the subject believes that there are raindrops on the window, but she does not believe that it is raining, nor that it is not raining.

Interpretability with an interpretation function does not suffer from the kind of indeterminacy that renders the simple account from Section 2.1 unusable. Before we interpret the subject we already need to specify which basic facts hold at which worlds of the set W to be able to encode the meaning of sentences with an interpretation function $I : \mathsf{At} \to \mathcal{P}W$ that maps into subsets of W. Any belief set that interprets the behavior of the subject with the interpretation I is then based a subset of W. Hence it is clear which basic facts obtain at which worlds in the interpreting belief set.

The account of interpretability with an interpretation function still leaves room for some indeterminacy. To see this consider again the example from above. The linguistic behavior in this example does not provide us with enough information to determine whether the subject believes that it is not raining. In the model given above the subject is uncertain to whether it is raining. But there is also another model with belief set $B' = \{v\}$ which generates the behavior $A = \{q\}$. And in this model the subject would believe that it is not raining.

In some cases, even when $\mathcal{V} = \mathcal{B}$, the indeterminacy can be quite extreme. For instance imagine that $I : \mathsf{At} \to \mathcal{P}W$ is such that every atomic sentences expresses the proposition W that is true at all worlds in the domain. It follows that every propositional sentence expresses relative to I either the whole domain or the empty set. Now assume that the subject accepts all sentences which are logical consequences of the set of all atomic sentences. Any non-empty belief set $B \subseteq W$ can then be extended to a model (W, B, I) that generates the behavior of the subject. Since these examples is constructed such that the subject is using a particularly poor language I do not take them to show that the account developed here does not fulfill the determinacy requirement.

To evaluate the account on the little-input and the variety requirement we first discuss the representation result. It is given in Proposition 7.1.3 and states that a behavior is interpretable with an interpretation if and only if it is closed under implication relative to the interpretation, which is defined as follows:

2.2.2. Definition. A behavior $A \subseteq \mathcal{V}$ is *closed under implication* relative to an interpretation function $I : \mathsf{At} \to \mathcal{P}W$ if for any sentence $\psi \in \mathcal{V}$ and set of sentences $F \subseteq \mathcal{V}$ we have that:

$$\text{If } \bigcap_{\varphi \in F} I(\varphi) \subseteq I(\psi) \text{ and } F \subseteq A \text{ then } \psi \in A.$$

We can think of this condition as saying that the subject needs to accept every sentences that is implied by the sentences that she already accepts given our prior assumption about the meaning of these sentences.

In the case where $\mathcal{V} = \mathcal{B}$, closure under implication entails that A is a theory of propositional logic. For, if ψ is a logical consequence of the sentences in A then it is implied by some $F \subseteq A$ relative to every interpretation I.

Depending on how we think about meanings in the language of the subject the account of interpretation given in this section has difficulties with either the little-input or the variety requirement. I sketch this problem in the remainder of this section.

Let us consider two cases. In the first case we think of facts about the meanings in the language of the subject as depending only on her linguistic behavior. I will argue that in this case the account of this section does not fulfill the little-input requirement. In the other case we think of facts about meanings as being independent of the subject's linguistic behavior. For this case I argue that we run into difficulties with the variety requirement. This case distinction is not exhaustive since it is also possible that facts about meanings depend on both the subject's behavior and states of the world that are independent from the subject's behavior. I would expect that with such a hybrid approach the account of interpretation of this section might run into similar difficulties with both the little-input and the variety requirement.

First consider the case where we take facts about meanings in the language of the subject to depend only on the subject's behavior. In this case it is presumably the task of an account of interpretation to determine these meanings from the behavior of the subject. Assuming that this information is already available to the interpreter in advance to interpretation violates the little-input requirement.

If we think about facts about meaning as depending on the behavior of the subject then the account has no difficulties with the variety requirement, other than the condition that the subject's behavior is a theory. The reason is that if the subject's behavior is not closed under implication relative to our hypothesis about meanings in the language then this would be evidence against this hypothesis. We would have to come up with a different hypothesis about meanings in the language of the subject that allows us to interpret the behavior of the subject.

In the other case we think about meanings in the language of the subject as being determined by some facts which are not depending on her linguistic behavior. These facts might be for instance be about the causal chains connecting the subject's utterance of certain expression to things in the world or facts about the usage of the term in the subject's community which are largely independent of the subject's own usage. Since such facts are independent of the subject own linguistic behavior it seems reasonable to assume that we could know them in advance to the interpretation of the subject's behavior. There would still be the difficulty of determining what the sentences in the language of the subject mean but that problem would be a different from the problem of interpreting the behavior of subject. Hence in this case the assumptions that we know the meaning of expressions in the language of the subject is not a problem for the little-input requirement.

If facts about meanings are determined independently of the subject's own linguistic behavior then we can no longer take a violation of closure under implication as showing that our hypothesis about the meaning of sentences in the subject's language is wrong. If it is plausible that there are behaviors which do not satisfy closure under implication with respect to the interpretation function that encodes the facts about meanings in the subject's language then this shows that the account given here does not satisfy the variety requirement.

It seems quite plausible that the subject's behavior might not satisfy closure under implication with respect to an interpretation that describes meanings in her language which are independent of her behavior. This happens for instance if the subject does not know whether the facts that determine the meaning of her sentences obtain, or if she thinks that the meaning of her sentences is determined by different facts. Every concrete example of such a linguistic behavior is going to be controversial since it hinges on what facts we take to determine the meaning of sentences in the language of the subject.

One example could be constructed from the sentence "The morning star is the evening star." Let us assume that the facts determining the meaning of this sentence are such that the expressions "morning star" and "evening star" refer to the same object in all possible worlds. Then this sentences expresses the proposition represented by the set of all worlds. Considering the instance of Definition 2.2.2 where $F = \emptyset$ this requires that the sentence "The morning star is the evening star" should be contained in every linguistic behavior. But we can easily imagine a subject which does not accept that sentence because she does not know that the morning star is the evening star.

Due to these problems with the assumption that we know in advance to interpretation what the sentences in the language of the subject mean I am not pursuing the approach of this section any further. From now on I am supposing that meanings in the language of the subject are only constrained by the interpretation of her behavior. Hence I am neglecting the possibility that these meanings might also depend on facts that are independent of her linguistic behavior.

References to the literature

The account of this section holds the meaning of sentences in the language of the subject constant and only concerns the problem of determining her beliefs. Because we assume the meaning of sentences to be know in advance and hence every sentence is associated with a fixed proposition we might also just ignore the distinction between sentences and propositions. I take this to be the approach that is usually taken in the work on doxastic logic.

2.3 Fixing beliefs

In this section I discuss an account of interpretation in which we assume to know the beliefs of the subject about the relevant basic facts prior to interpretation. The beliefs of the subject are represented by a belief set $B \subseteq W$, which is a subset of some fixed set W of possible worlds. Because in this way the domain W can be fixed by us in advance to interpretation this account is not prone to the indeterminacy discussed at the end of Section 2.1. The account of this section still faces two difficulties which I discuss in this section and solve in the following three sections.

As in the previous two sections a linguistic behavior is again a set of atomic or propositional sentences. Also the linguistic behavior generated by a model is again the set of all sentences that express proposition that is true at all worlds in the belief set of the model. What is different in the account of this section is the notion of interpretability.

We are assuming that we the interpreters know in advance to interpretation all the beliefs that the subject has about the basic facts. To represent this assumption formally we construct a domain of relevant worlds W that contains all possible combination of basic facts as worlds. A belief about these relevant basic facts is then represented by a subset of W. Since we take the beliefs of the subject to be closed under implication of propositions we can represent the set of all beliefs that we assume the subject to have by just one belief set $B \subseteq W$ that is the intersection of all of her beliefs.

We want to interpret the subject such that according to the interpreting simple possible world model she has the belief set B that encodes our prior knowledge about her beliefs. The only thing that interpretation has to determine is the interpretation function that encodes the language of the subject. This yields the following notion of interpretability:

2.3.1. Definition. An interpretation function $I : \mathsf{At} \to \mathcal{P}W$ *tightly interprets a linguistic behavior* $A \subseteq \mathcal{V}$ *with a belief set* $B \subseteq W$ if $A = A^M$ for the simple possible world model $M = (W, B, I)$.

A linguistic behavior $A \subseteq \mathcal{V}$ is *tightly interpretable with a belief set* $B \subseteq W$ if there is some interpretation function $I : \mathsf{At} \to \mathcal{P}W$ that interprets A with B.

I call the notion defined here tight interpretability to distinguish it from a similar notion of interpretability introduced in Section 2.4 on which there is a looser connection between the belief set that captures our prior knowledge about the subject and the belief set of an interpreting model.

Let us consider an example which shows how the notion of tight interpretability with some belief set avoids the indeterminacy from the end of Section 2.1. Imagine that we observe the subject in a situation where it is raining heavily. The rain is falling onto her face and her coat is getting wet from the water. It seems reasonable to assume that in such a situation the subject believes that it is raining. We can model this relative to the domain $W = \{w, v, u\}$ that we have been using before. That is, w is a world at which it is raining and the there are raindrops on the window, at v it is not raining but there are still raindrops on the window and at u it is not raining and the raindrops have dried up. Our assumptions that the subject believes that it is raining means that her belief set must only contain the world w since this is the only world in the domain where it is raining. Moreover, we might assume that she is not having contradictory beliefs which means that her belief set is not empty. Putting this together we model this situation as one in which we the subject's belief set is $B = \{w\}$.

Assume that in this situation we only care about the meaning of the atomic sentence p. Hence $\mathcal{V} = \mathsf{At} = \{p\}$. Moreover, let us suppose that we find that the subject accepts p and all propositional consequences thereof. So her linguistic behavior is the set $A = \mathsf{cl}\,(\{p\})$.

The behavior A is tightly interpretable with $B = \{w\}$. An interpretation function $I : \mathsf{At} \to \mathcal{P}W$ which interprets A with B is defined such that $I(p) = \{w\}$.

In this example the problem from Section 2.1 does not arise. We have already specified in advance what basic facts obtain at which worlds from the domain W. Hence we know that according to $M = (W, B, I)$ the sentence p means that it is raining because according to I the sentence p is true exactly at those worlds where it is raining. We also know that according to M the subject believes that it is raining because all her doxastic alternatives in B are worlds where it is raining.

The account of this section satisfies the determinacy constraint in the sense that given a model which interprets the subject we are able to tell what the subject believes and what her sentences mean according to that model. There is however a weaker form of indeterminacy left that is still problematic. For many linguistic behaviors and fixed belief sets there are many distinct interpretation functions that interpret that behavior with the belief set.

For instance in the example from above we could also tightly interpret the behavior of the subject with the belief set $\{w\}$ using a model based on an interpretation with $I(p) = \{w, u, v\}$. According to such a model the sentence p just expresses a tautology that is always true. This is a different meaning than in the model given above, where p means that it is raining. This shows that the account given in this section does not determine whether p means that it is raining or expresses a tautology. The sentence p might express any proposition that is a superset of the belief set.

This problem is very general. For every sentence that the subject accepts we can conclude that this sentence is true in all worlds that she considers possible. But we do not know whether that sentence is true or false at any world that she does not consider possible. The meaning of sentences that the subject does not accepts is even more indeterminate. In this case we can only conclude that she considers a world possible where the sentence is false. If she considers multiple worlds possible it is not determined which world this is and for any other world it is not determined whether the sentence is true or false there.

To reduce this indeterminacy we might try to further constrain the possible meanings of sentences in the language of the subject by observing the subject's acceptance of these sentences across many different situations. For instance imagine that we later observe the subject from the example above sunbathing on the beach. In this situation she clearly believes that it is not raining and let us suppose that she is uncertain to whether there are raindrops on some window far away. In this situation her belief set is $\{u, v\}$. If in this situation the subject does no longer accept the sentence p this tells us that p can not be a tautology. It then either means $\{w\}$, $\{w, u\}$ or $\{w, v\}$. If we find further that the subject accepts the sentence $\neg p$ when she is at the beach then we know that the meaning of p must be the set $\{w\}$. Otherwise, we have to look at yet other situations where the subject has different beliefs to further constrain the possible meanings of p. In Sections 2.5 and 2.6, I implement the idea of considering the subject's behavior across multiple situations as a formal account of interpretation.

There is however another problem that I want to address before I extend the account given here to multiple situations. The problem arises if we try to prove a representation result for tight interpretability with a belief set.

In the case where $\mathcal{V} = \mathsf{At}$, the representation result is given by Corollary 7.3.3 which follows from more complex results for the setting of Section 2.5. Every behavior is tightly interpretable with some fixed non-empty $B \subseteq W$ and only the behavior $A = \mathsf{At}$ is tightly interpretable with the empty set $B = \emptyset$. This is quite intuitive since we can choose the meaning of the sentences in At such that it is either true or false at all of the worlds in B depending on whether the subject accepts the sentence.

In the case where $\mathcal{V} = \mathcal{B}$, it is surprisingly difficult to obtain a representation result for tight interpretability. The result is given by Corollary 7.3.11. To state the conditions on tight interpretability with a belief set $B \subseteq W$ we need to distinguish cases depending on whether the belief set B is infinite or finite. For the infinite case we need the additional assumption that set of atomic sentences At is at most countably infinite. We can then show that a behavior $A \subseteq \mathcal{B}$ is tightly interpretable with some infinite belief set $B \subseteq W$ if and only if A is a consistent theory. In the finite case we have that a behavior $A \subseteq \mathcal{B}$ is tightly interpretable with some finite set $B \subseteq W$ containing n elements if and only if A is an n-theory. This result involves the notion of an n-theory that is introduced in Definition 7.3.6 in Chapter 7. For the discussion here it is sufficient to know that an n-theory is a theory that satisfies an additional condition which depends on the number n.

To get an idea why it is not sufficient for tight interpretability to just require that A is a theory consider the case where $B \subseteq W$ is a singleton set $B = \{w\}$. In this case the representation result says that the subject's linguistic behavior is interpretable only if it is a 1-theory. By instantiating Definition 7.3.6 we see that a 1-theory is same as a complete theory. It follows that to be tightly interpretable with the singleton set $\{w\}$ the subject needs to accept either φ or $\neg\varphi$ for every sentence $\varphi \in \mathcal{V}$. One can easily see that this constraint follows from the fact that $\{w\}$ contains only one world. No matter how we define the interpretation $I : \mathsf{At} \to \mathcal{P}W$ it is the case that for every sentence φ either $w \in I(\varphi)$ or that $w \in I(\neg\varphi)$. It follows that for every sentence that either the sentence or its negation is true at every world in the belief set $\{w\}$. Hence, the subject accepts either the sentence or its negation. If a set of sentences does not have this property then it can not be the linguistic behavior generated by a model in which belief set contains exactly one world.

I have two reasons for not being satisfied with the condition that the set of sentences that the subject accepts is an n-theory whenever her belief set contains a finite number of n elements.

The first is a practical problem. The notion of an n-theory is technically rather involved and hence becomes quite a nuisance when we attempt to prove representation results for the more complex settings discussed in the following chapters. It is convenient to get this notion out of the way as soon as possible.

The second reason is of actual conceptual relevance. One can argue that an account that requires that the subject's behavior is an n-theory whenever the belief set B contains a finite number of n elements does not fulfill the variety requirement. To do so we can consider cases in which the language of the subject contains vague expressions.

To give a concrete example assume that the subject happens to be a speaker of English and so her language contains the sentence "The man is tall." that we abbreviate with p. Let us consider a domain $W = \{w_{1.6}, w_{1.8}, w_{2.0}\}$ with three possible worlds which are such that at w_h the man that is the sentence p is about is h meters tall. Now let us suppose that we are in a situation where there is a man 1.8 meters tall next to the subject. If we also assume that the subject notices this then it follows that her belief set is the singleton $B = \{w_{1.8}\}$. Because a man of 1.8 meters height is presumably a borderline case for the application of the predicate "tall" in English it seems reasonable to expect that the subject does neither accept the sentence p nor the sentence $\neg p$. Hence her linguistic behavior is not a complete theory even though her belief set contains just one world.

In this next section I give simple modification of the notion of interpretability with a belief set B that avoids the condition that the subject's behavior is an n-theory whenever B contains n elements.

2.4 Splitting worlds

In this section I introduce the notion of splitting interpretability with a belief set $B \subseteq W$ that overcomes the condition on tight interpretability with B that requires that the subject's behavior is a n-theory whenever B contains a finite number of n elements. The notion of splitting interpretability lets us duplicate worlds in the domain W when we are interpreting the subject.

The reason for the strong constraints on tight interpretability with B is that by fixing the belief set B we do not just fix the strongest proposition that the subject believes but also how many worlds are available to account for her remaining uncertainty. Splitting possible worlds allows us to avoid the latter without giving up the former.

Let me explain on an example how such splittings of worlds are supposed to work. Consider again the situation where the subject knows that there is a man next to her that is 1.8 meters tall. Her belief set contains exactly the world $w_{1.8}$ from the set of worlds $W = \{w_{1.6}, w_{1.8}, w_{2.0}\}$, where at w_h the man is h meters tall. Let p be the English sentence "The man is tall." The problem is that with this belief set $\{w_{1.8}\}$ the subject must accepts either the sentence p or its negation $\neg p$ since either $w_{1.8} \in I(p)$ or $w_{1.8} \notin I(p)$ for any interpretation I.

We could avoid this conclusion if we had another copy $w'_{1.8}$ of $w_{1.8}$ in which the man is also 1.8 meters tall. The subject, who knows that the man is 1.8 meters tall, could then have the belief set $\{w_{1.8}, w'_{1.8}\}$. If we now define the interpretation I such that $I(p) = \{w'_{1.8}, w_{2.0}\}$ it would turn out that the subject accepts neither p nor $\neg p$.

Even though this splitting of the world $w_{1.8}$ is technically helpful it is not easy to give an intuitive explanation of what it is doing. We think of $w'_{1.8}$ as a copy of $w_{1.8}$ insofar that the same basic facts obtain at both worlds. For this reason we can copy worlds in the subject's belief set without changing her beliefs about these basic facts.

But if the same basic facts obtain at both $w_{1.8}$ and $w'_{1.8}$ how can the sentence p be true at $w_{1.8}$ and not at $w'_{1.8}$? A simple explanation is that even though the two worlds are the same for all the basic facts they are distinct with respect to semantic facts that determine the meaning of p. At $w_{1.8}$ the semantic facts are such that "tall" applies to men that are 1.8 meters tall, whereas at $w'_{1.8}$ the semantic facts are such that "tall" does not apply to men of this height.

This explanation of the difference between $w_{1.8}$ and $w'_{1.8}$ hinges on a particular view of the relation between belief and meaning that many readers might not find intuitive. I am going to develop this view more carefully in Chapter 5, after I have said much more about meaning in Chapter 4. In Chapter 5 I also compare splitting interpretability to another notion of interpretability that does not split possible worlds but still lifts the condition that the behavior of the subject is a n-theory whenever her belief set contains n elements.

To make the process of copying possible worlds formally precise I use the notion of a splitting of a domain of possible worlds.

2.4.1. Definition. A *splitting of a domain* W is a set W' of *split worlds* together with a surjective function $f : W' \to W$, which is called *splitting function.*

We think of the domain W' in a splitting $f : W' \to W$ of W to be the set of all worlds in W plus all the needed copies of these worlds. The surjective function f maps any world in W' to a world in W that it is a copy of. We think of a world

$w' \in W'$ as corresponding to the same combination of basic facts as the world $f(w') \in W$ in the original domain W.

The splitting corresponding to the example above is given by the function $f : W' \to W$ where $W' = \{w_{1.6}, w_{1.8}, w'_{1.8}, w_{2.0}\}$, $W = \{w_{1.6}, w_{1.8}, w_{2.0}\}$ and $f(w_{1.6}) = w_{1.6}$, $f(w_{1.8}) = f(w'_{1.8}) = w_{1.8}$ and $f(w_{2.0}) = w_{2.0}$.

Every proposition $P \subseteq W$ corresponds to a proposition

$$P' = f^{-1}[P] = \{w' \in W' \mid f(w) \in P\}$$

over the new set W' of split worlds. The proposition $f^{-1}[P]$ contains all the split worlds where the same basic facts obtain as at one world in P. In the example we have for instance that the proposition $\{w_{1.8}\}$ over W corresponds to the proposition $f^{-1}[\{w_{1.8}\}] = \{w_{1.8}, w'_{1.8}\}$ over W'.

In the definition of splitting interpretability with a belief set we want to allow the domain of an interpreting model to be a splitting of the original domain W relative to which the belief set is defined. The process of interpretation should determine an interpretation function over some splitting of the given domain W. This information, a splitting together with an interpretation function over the split set of worlds, can be put into a single structure.

2.4.2. Definition. A *splitting interpretation model* over W is triple (W', f, I), where W' is a splitting of W with the surjective splitting function $f : W' \to W$ and $I : \mathsf{At} \to \mathcal{P}W'$ is an interpretation function that maps atomic sentences to propositions over the split set of worlds W'.

There is one additional obstacle to defining when a splitting interpretation model (W', f, I) interprets a linguistic behavior with a belief set $B \subseteq W$. The belief set B is defined over the domain W, but the interpretation $I : \mathsf{At} \to \mathcal{P}W'$ in the splitting interpretation model is defined with respect to the split domain W' which is in general distinct from W. Hence these two objects do not fit together into one simple possible world model from which we can generate the behavior of the subject. The solution to this problem is to turn the belief set B, which is defined relative to W, into a belief set B' relative to W'. This belief set is defined as $B' = f^{-1}[B]$, where $f : W' \to W$ is the splitting function. With this definition of B' it is ensured that the subject believes a proposition $P \subseteq W$ about the basic fact in W if and only if she believes the corresponding proposition $f^{-1}[P] \subseteq W'$ over the split set of worlds.

We now obtain the following definition of splitting interpretability:

2.4.3. Definition. A splitting interpretation model (W', f, I) *interprets a linguistic behavior $A \subseteq \mathcal{V}$ with a belief set $B \subseteq W$* if $A = A^M$ for the simple possible world model $M = (W', f^{-1}[B], I)$.

A behavior $A \subseteq \mathcal{V}$ is *splitting interpretable with a belief set $B \subseteq W$* if there is some splitting interpretation model (W', f, I) that interprets A with B.

An example of a splitting interpretation model is the triple (W', f, I) where W', f, and I are like in the examples discussed throughout this section. This model interprets the behavior $A = \mathsf{cl}\,(\emptyset)$, which contains neither p nor $\neg p$, with the belief set $B = \{w_{1.8}\} \subseteq W$.

In Corollary 7.3.5 it is shown that a behavior $A \subseteq \mathcal{B}$ is splitting interpretable with a non-empty $B \subseteq W$ if and only if it is a consistent theory. It is splitting interpretable with the empty set $\emptyset \subseteq W$ if and only if it is the inconsistent theory.

The notion of splitting interpretability with a belief set avoids the difficulty that the number of elements in a belief set constrains the theories which are tightly interpretable with the belief set. We are trading a more complex notion of interpretability for a simpler representation results. This is crucial to still obtain succinct condition on interpretability for the more complex settings treated later in this thesis.

I end this section with three remarks that should clarify the theoretical purpose of splitting worlds.

First, splitting interpretability does not suffer from the kind of indeterminacy that is observed at the end of Section 2.1. One might fear that this is the case because the interpreting interpretation function is defined over a domain W' that might be distinct from the domain W that has been fixed in advance and for which we know at what worlds what basic facts obtain. The splitting function $f : W' \to W$ tells us how to think of the worlds in W'. A basic fact obtains at a world $w' \in W$ if and only if it obtains at $f(w') \in W$, where the latter has been specified prior to interpretation. Hence there is no indeterminacy about what basic facts obtain at the worlds in W'.

Second, the splitting of worlds is not a tool to capture the indeterminacy of interpretation. It is possible to describe cases of vagueness as cases in which the meaning of some sentence in the language of the subject is indeterminate. For instance it is indeterminate whether the sentence "The man is tall." is true at a world where the man is 1.8 meters tall. But this is not the same kind of indeterminacy as the indeterminacy of interpretation that bothers us when the behavior of the subject does not determine a unique interpreting model. In cases of vagueness the behavior of the subject completely determines that the meaning of some sentences is indeterminate.

Third, the splitting of worlds should not be used in cases where the linguistic behavior of the subject is sensitive to facts that we the interpreters have failed to include in the relevant basic facts when specifying the original domain W. In such situations the behavior of the subject might be similar as in cases of vagueness in that it violates the condition that it is an n-theory, whenever her belief set contains n worlds. Nevertheless, such cases should not be accounted for by a splitting of worlds. Instead, we should use a different domain than W that includes all the facts that the subject's behavior is sensitive to as basic facts.

Let me explain this third point more extensively on an example. Assume again that we fix the domain $W = \{w_{1.6}, w_{1.8}, w_{2.0}\}$ prior to interpretation. The relevant basic facts are about the height of some man and they are such that at w_h the man is h meters tall. We interpret the subject in some situation s where she has the information that the man is 1.8 meters tall. So we have that relative to the domain W her belief set is $B = \{w_{1.8}\}$. The subject's behavior is such that in s she does not accept p nor does she accept $\neg p$. Because A is not a complete theory even though B is a singleton set it is not possible to tightly interpret this behavior. Follow the reasoning from this section we might conclude that this is a case of vagueness that can be accounted for by the notion of splitting interpretability.

Imagine, however, that actually there is a different explanation for the behavior of the subject that is unknown to us the interpreters. She is using the sentence p to express that it is raining. In the situation s she does accept neither p nor $\neg p$ because she is uncertain whether it is raining. We are unable to tightly interpret the subject with $B \subseteq W$ not because of vagueness but because we have failed to include facts about the weather into the domain W.

The problem is that we are starting from a wrong hypothesis about which facts are relevant for specifying the belief function of the subject. It would be wrong to solve this problem by simply interpreting the subject with a splitting over W, as for instance the splitting in the example further above where p is the vague English sentence "This man is tall." Such a splitting interpretation function does not tell us that p means that it is raining. The right solution to the problem would be to construct a different domain that includes basic facts about the weather. We should for instance use the domain $\{w_{1.6}, w_{1.8}, w_{2.0}, v_{1.6}, v_{1.8}, v_{2.0}\}$ such that for every h it is raining at the world w_h, not raining at the world v_h, and at both worlds the man is h meters tall. Relative to this domain we have that $B = \{w_{1.8}, v_{1.8}\}$ and hence the behavior of the subject is tightly interpretable.

This example raises the question how we can distinguish between cases in which the subject is using vague expressions and we should interpret her behavior with a splitting interpretation model and cases in which we have failed to include a relevant basic facts when constructing the domain and hence should revise the domain with which we are interpreting the subject. I think that this is an important problem but I do not know a good solution.

2.5 Multiple situations

In this section I formalize the idea of interpreting the behavior of the subject across different situations. As we see in the following section this resolves the indeterminacy of meanings that is observed in Section 2.3.

To give a formal account of how we interpret a subject's behavior in multiple situations we need to formally represent these situations in our model. For this purpose we fix some set S, whose elements we call *situations*.

Let us consider an example that is already suggested in Section 2.3. We imagine that we are observing the subject in two different situations. In the first situation, which we call s_R, the subject is standing in the rain and getting wet. In the second situation, which we call s_B, the subject is sunbathing on the beach. Assuming that these are all the situations in which we have observe the subject we can model the example with the set of situations $S = \{s_R, s_B\}$.

Situations are just the elements of the set S. They do not have any internal structure. To associate situations with information that is relevant for interpretation I am using functions which map situations to the formal representations of this relevant information.

A first kind of information that is relevant for interpretation are the sets of sentences that the subject accepts. In different situations the subject might accept different sentences. This yields the following definition of a linguistic behavior:

2.5.1. Definition. A *linguistic behavior* is a function $a : S \to \mathcal{PV}$ which maps a situation $s \in S$ to the set of sentences $a(s) \subseteq \mathcal{V}$.

We think of the linguistic behavior $a : S \to \mathcal{PV}$ to be defined such that for any sentence $\varphi \in \mathcal{V}$ we have that $\varphi \in a(s)$ if and only if the subject accepts the sentence φ in the situation s.

The terminology from this Definition 2.5.1 conflicts with the terminology introduce in Definition 2.1.1 where a linguistic behavior is defined to be a set $A \subseteq \mathcal{V}$. This should not cause any problems since from the context it is always clear which notion is meant.

As an example I use again the setting suggested at the end of Section 2.3. We interpret the subject in the set $S = \{s_R, s_B\}$ of situations that is discussed above. In s_R she is standing in the rain, and in s_B she is sunbathing on the beach. Let us assume that in this example the subject's language contains all the propositional formulas generated from just one propositional letter p. Now we observe that as the subject is standing in the rain she is accepting all the propositional consequences of the sentence p. And when she is sunbathing on the beach, she accepts all the propositional consequences of the sentence $\neg p$. This linguistic behavior can be represented by the function $a : S \to \mathcal{PB}$ where $a(s_R) = \mathsf{cl}(\{p\})$ and $a(s_B) = \mathsf{cl}(\{\neg p\})$.

I now discuss what kind of possible world models can be used to generate a linguistic behavior as defined in Definition 2.5.1. These models should be similar to the simple possible world models from Section 1.2 in that they contain a formal representation of the subject's beliefs and of the meaning of sentences in her language.

The point of the account in this section is that the subject can have different beliefs in different situations. This allows us to narrow down the possible meanings of the sentence by observing how her acceptance of the sentence changes while varying her belief set. Hence we need to make the notion of a belief set dependent on the situation in which we are interpreting the subject.

2.5.2. Definition. A *belief function* is a function $b : S \to \mathcal{P}W$ which maps a situation s to the set of worlds $b(s) \subseteq W$.

As an illustration consider again the example where we have the two situations $S = \{s_R, s_B\}$. We represent the beliefs of the subject relative to a domain $W = \{w, v, u\}$, where w is the only rainy world, and in w and v but not in u there are raindrops on the window. Let us make the assumption that in the situation s_R, where the subject is standing in the rain, she believes that it is raining. In the other situation s_B the subject is laying on the beach in the sun. We assume that in this situation the subject believes that it is not raining but she has no particular beliefs to whether there are still some raindrops on the window of some building far away. We can encode this assumptions with the belief function $b : S \to \mathcal{P}W$ where $b(s_R) = \{w\}$ contains only the rainy world w and $b(s_B) = \{v, u\}$ contains the sunny worlds v and u which leaves it open whether there are raindrops on the window.

In this section I represent the meaning of sentences in the language of the subject by an interpretation function that is independent of the situation in which we are interpreting the subject. Hence we are assuming that the propositions that are expressed by sentences in the language of the subject can not change across different situations. This assumption is motivated by our original reason for considering multiple situations. If the subject's interpretation function was to change across situations we would again end up with the indeterminacy, which has troubled us in Section 2.3, that acceptance of sentence one single situation is not enough to determine an interpretation function.

It might seem at this point that assuming that the subject's interpretation function is the same in every situation amounts to assuming that the meaning of sentences in the subject's language can not change. This is suggested by my terminology of calling the proposition that a sentence expresses according to an interpretation function the meaning of the sentence in the language of the subject. This terminology is however somewhat simplistic. In Chapter 4 I introduce a more complex representation of meanings that allows for a different view on the interpretation function as it is used in the setting of this section and with which it is not adequate to say that meanings in the language of the subject can not change. I return to such terminological matters in Section 8.2 at the end of this thesis.

The discussion in the preceding paragraphs leads to the following notion of possible world models:

2.5.3. Definition. A *multi-situation model* M is a triple $M = (W, b, I)$, with a domain W of possible worlds, a belief function $b : S \to \mathcal{P}W$ and an interpretation function $I : \mathsf{At} \to \mathcal{P}W$.

Note that this definition depends on the fixed sets S and At. These sets need to be chosen such that they match the corresponding sets in the type of the linguistic behaviors $a : S \to \mathcal{P}\mathcal{V}$ that we want to interpret.

To define the linguistic behavior generated by some multi-situation model we formalize the acceptance principle which states that the subject accepts a sentence if and only if she believes the proposition that the sentences expresses.

2.5.4. Definition. The *linguistic behavior* $a^M : S \to \mathcal{PV}$ *generated* by a multi-situation model $M = (W, b, I)$ is defined such that for all situations $s \in S$

$$a^M(s) = \{\varphi \in \mathcal{V} \mid b(s) \subseteq I(\varphi)\}.$$

This definition is an extension of Definition 2.1.2 to multiple situations. Given a multi-situation model $M = (W, b, I)$ one can define the simple possible world model $M_s = (W, b(s), I)$ for any situation $s \in S$. This model M_s models the beliefs of the subject in the situation $s \in S$ according to the model M. The above definition of the behavior $a^M : S \to \mathcal{PV}$ generated by a multi-situation model M extends the Definition 2.1.2 of the behavior $A^{M'} \subseteq \mathcal{V}$ generated by a simple possible world model $M' = (W, B, I)$ in the sense that $a^M(s) = A^{Ms}$ for every situation $s \in S$.

Again, consider our example where $S = \{s_R, s_B\}$, $W = \{w, v, u\}$, $b(s_R) = \{w\}$, and $b(s_B) = \{v, u\}$. The linguistic behavior $a : S \to \mathcal{PB}$ from the beginning of this section where $a(s_R) = \mathsf{cl}(\{p\})$ and $a(s_B) = \mathsf{cl}(\{\neg p\})$ is the behavior generated by the model $M = (W, b, I)$ where $I(p) = \{w\}$.

I conclude this section by introducing a notion that only becomes relevant later in Section 4.2 but that might help to clarify the relation between situations and possible worlds. For every situation $s \in S$ there is some world $w_s \in W$ that describes all the facts that actually obtain in s. This world w_s is the actual world of the situation s. The actual world has to be relativized to situations because different basic facts might obtain in different situations.

As an example consider again the set of situations $S = \{s_R, s_B\}$ and the set of worlds $W = \{w, v, u\}$ that are used in the examples above. In the situation s_R, where the subject is standing in the rain, we have that $w_{s_R} = w$ because in this situation it is raining and there are raindrops on the window. In the other situation s_B the subject is sunbathing on the beach. It is not raining in s_B and we might also stipulate that there are no raindrops on the window. Hence $w_{s_B} = u$.

References to the literature

For readers who are familiar with two-dimensional semantics it might seem that situations are similar to contexts as they are used for instance by Kaplan (1989) to capture the context-dependence of meaning. Both, situations, as used above, and contexts, as used in formal semantics, are meant to formally represent a setting in which a subject or speaker is accepting or uttering sentences. I do not know whether it is appropriate to identify situations with contexts.

There is however a good reason for not calling situations contexts. Later in this thesis I employ the formal framework of two-dimensional modal logic to represent variance of semantic facts across possible worlds. Two-dimensional modal logic is also used in formal semantics to account for the context-dependence of meaning and in this application contexts play a particular role in the formal system. In my application of two-dimensional modal logic later in the thesis situations are not playing the role that is usually taken up by contexts. Hence it would become confusing if I was to call situations contexts.

2.6 The basic account

In this section I combine the ideas from the previous three sections into one single account of interpretation. The resulting account is the common base for the extensions developed in later chapters.

As in Sections 2.3 and 2.4 I make in this section the assumption that prior to interpretation we know what the subject believes about the relevant basic facts. To formally represent these beliefs we again fix a domain W of possible worlds such that for every relevant basic fact it is specified at which worlds in W it obtains. As in Section 2.5 we want to interpret the subject across different situations in which she can have different beliefs. For this we fix a set of situations S and use a belief function $b : S \to \mathcal{P}W$ to encode our prior knowledge about the beliefs of the subject. We are now looking for a good notion of interpretability with a given belief function $b : S \to \mathcal{P}W$.

A behavior is interpretable with a belief function $b : S \to \mathcal{P}W$ if it is the behavior generated by some model that is in some appropriate sense based on the belief function. The sense in which the model is based on the belief function depends on whether we choose tight or splitting interpretability.

Tight interpretability requires the domain of any model that interprets a behavior to be the set worlds that we have fixed in advance when specifying the belief function.

2.6.1. Definition. An interpretation function $I : \mathsf{At} \to \mathcal{P}W$ *interprets a linguistic behavior* $a : S \to \mathcal{PV}$ *with a belief function* $b : S \to \mathcal{P}W$ if $a = a^M$ for the model $M = (W, b, I)$.

A linguistic behavior $a : S \to \mathcal{PV}$ is *tightly interpretable with a belief function* $b : S \to \mathcal{P}W$ if there is some interpretation function $I : \mathsf{At} \to \mathcal{P}W$ that interprets a with b.

For splitting interpretability with a belief function $b : S \to \mathcal{P}W$ we allow that an interpreting interpretation function is defined over a splitting $f : W' \to W$ of W. For this purpose we use the notion of a splitting interpretation model from Definition 2.4.2. Similar to the single-situation case in Definition 2.4.3 we have the problem of lifting the given belief function $b : S \to \mathcal{P}W$ to a belief function $b' : S \to \mathcal{P}W'$ over the split domain. We want that in every situation $s \in S$ according to the lifted function b' the subject has the same belief about the basic facts encoded in W as she has according to b. The only way to guarantee this is to define $b' : S \to \mathcal{P}W'$ such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$. Thus we obtain the following definition.

2.6.2. Definition. A splitting interpretation model (W', f, I) *interprets a linguistic behavior* $a : S \to \mathcal{PV}$ *with a belief function* $b : S \to \mathcal{P}W$ if $a = a^M$ for the multi-situation model $M = (W', b', I)$ where $b' : S \to \mathcal{P}W'$ is defined such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$.

A linguistic behavior $a : S \to \mathcal{PB}$ is *splitting interpretable with a belief function* $b : S \to \mathcal{P}W$ if there is some splitting interpretation model (W', f, I) that interprets a with b.

The setting from the Sections 2.3 and 2.4 is recovered in the setting from this section as the case where $S = \{s\}$ is the singleton set. One finds that $a(s)$ is tightly, or respectively splitting, interpretable with $b(s)$ if and only if a is tightly, or respectively splitting, interpretable with b.

I continue by evaluation the account of interpretation from this section on the requirements from Section 1.4.

To evaluate the present account on the variety requirement it is helpful to have a look at the relevant representation results.

First consider the case where $\mathcal{V} = \mathsf{At}$ and hence we are not having any hypothesis about which parts of the subject's language correspond to the propositional connectives. For behaviors $a : S \to \mathcal{P}\mathsf{At}$ it is shown in Theorem 7.3.2 that tight and splitting interpretability with a belief function $b : S \to \mathcal{P}W$ are the same. So there is no need to consider the more complex notion of splitting interpretability in this case. Theorem 7.3.2 also shows that a behavior $a : S \to \mathcal{P}\mathsf{At}$ is tightly interpretable with a belief function $b : S \to \mathcal{P}W$ if and only if it satisfies the following condition:

2.6.3. Definition. A linguistic behavior $a : S \to \mathcal{PV}$ satisfies the *simple covering condition* relative to a belief function $b : S \to \mathcal{P}W$ if for all $s \in S$ and subsets $T \subseteq S$:

$$b(s) \subseteq \bigcup_{t \in T} b(t) \text{ implies } \bigcap_{t \in T} a(t) \subseteq a(s).$$

To discuss the atomic covering condition I first split it into two simpler conditions. The first condition is as follows:

2.6.4. Definition. A linguistic behavior $a : S \to \mathcal{PV}$ satisfies the *exact covering condition* relative to a belief function $b : S \to \mathcal{P}W$ if for all $s \in S$ and $T \subseteq S$:

$$b(s) = \bigcup_{t \in T} b(t) \text{ implies } \bigcap_{t \in T} a(t) \subseteq a(s).$$

The second condition is as follows:

2.6.5. Definition. A linguistic behavior $a : S \to \mathcal{P}\mathcal{V}$ satisfies the *monotonicity condition* relative to a belief function $b : S \to \mathcal{P}W$ if for all $s, t \in S$:

$$b(s) \subseteq b(t) \text{ implies } a(t) \subseteq a(s).$$

It is easy to see that the simple covering condition entails both the exact covering condition and the monotonicity condition. Conversely, it is shown in Proposition 7.2.1 that, if the set $b[S] = \{b(s) \mid s \in S\}$ is closed under non-empty unions, then the conjunction of the exact covering condition and the monotonicity condition entails the simple covering condition. That $b[S]$ is closed under unions means that for every set of situations $T \subseteq S$ there is a situation $d \in S$ such that $b(d) = \bigcup\{b(t) \mid t \in T\}$. It is not plausible that the set $b[S]$, which is the set of all belief sets with which we are interpreting the subject in some situation, is closed under non-empty unions. Nevertheless I suppose that the conjunction of the exact covering condition and the monotonicity condition gives us an idea of the strength of the simple covering condition.

The exact covering condition roughly requires that if a subject accepts some sentence in all of an exhaustive set of possible ways of refining her beliefs about the world then she accepts the sentence independently of one of these refinements. I take this to be a quite plausible constraint.

Let me illustrate the exact covering condition with an example. Assume we are observing the subject in three different situations s_B, s_D and s_U. In s_B she is lying in the sun on the beach. She believes that it is not raining but she does not have any beliefs as to whether there are raindrops on the window of some building far away. We can make this formal using the domain $W = \{w, v, u\}$ such that in w but not in v and u it is raining and at w and v but not u there are raindrops on the window. Relative to this domain the belief function of the subject is such that $b(s_B) = \{v, u\}$. We also observe the subject in situation s_D, where she is standing next to the window of the building and it has just stopped raining but there are still raindrops on the window. Her belief set in this situation is $b(s_D) = \{v\}$. In the last situation s_U the subject is standing next to the window and the raindrops have dried up. Thus her belief set is the set $b(s_U) = \{u\}$. The important point about this example is that in s_D and s_U the subject has stronger beliefs about the world than in s_B and that s_D and s_U together cover all possible ways in which these beliefs might be stronger. Now suppose that there is some sentence p such that $p \in a(s_D)$ and $p \in a(s_U)$. This says that the subject accepts p when she believes that it is sunny and there are no raindrops on the window and she accepts p when she believes that it is sunny and there are raindrops on the window. Given her belief that it is sunny, her acceptance of p is independent of her belief about whether there are raindrops on the window. The exact covering condition requires then that $p \in a(s_B)$, that is, the subject also accepts p if she just believes that it is sunny and is uncertain whether there are raindrops on the window.

The monotonicity condition requires that if in some situation the subject believes everything that she also believes in some other situation then she also accepts every sentence that she accepts in the other situation. This is an obvious constraint that follows directly from our approach to link acceptance to belief. It seems that a counterexample to the monotonicity condition would rather show that our assumptions about the subject's belief sets in the different situation are wrong than that it would show that the account of interpretation of this section is not rich enough.

Because I take both the exact covering condition and the monotonicity condition to be a plausible constraint on linguistic behaviors and they are roughly equivalent to the simple covering condition I also take the simple covering condition to be a plausible constraint. Hence tight interpretability with a belief function seems to satisfy the variety requirement in the case where $\mathcal{V} = \mathsf{At}$.

Next consider the case where $\mathcal{V} = \mathcal{B}$, that is, we assume to know which constructions in the language of the subject correspond to the propositional connectives. In this case there is a difference between tight and splitting interpretability.

I am only able to give a general characterization of the interpretable behaviors in the case of splitting interpretability. Theorem 7.3.4 states that a behavior $b : S \to \mathcal{PB}$ is splitting interpretable with a belief function $b \subseteq S \times W$ if and only if it satisfies two conditions. The first is as follows:

2.6.6. Definition. A behavior $a : S \to \mathcal{PB}$ satisfies the *conjunctive covering condition* relative to the belief function $b : S \to \mathcal{P}W$ if for all $s \in S$ and $s_{j,k} \in S$ for every $j \in J$, of some index set J, and $k \in \{1, \dots, n_j\}$, for some number n_j, it holds that

$$b(s) \subseteq \bigcup_{j \in J} (b(s_{j,1}) \cap \cdots \cap b(s_{j,n_j})) \text{ implies } \bigcap_{j \in J} \mathsf{cl}\left(a(s_{j,1}) \cup \cdots \cup a(s_{j,n_j})\right) \subseteq a(s).$$

The second condition is as follows:

2.6.7. Definition. A behavior $a : S \to \mathcal{PB}$ satisfies the *conjunctive consistency condition* relative to the belief function $b : S \to \mathcal{P}W$ if for all $s_1, \dots, s_n \in S$

$$b(s_1) \cap \cdots \cap b(s_n) \neq \emptyset \text{ implies that } \mathsf{cl}\left(a(s_1) \cup \cdots \cup a(s_n)\right) \text{ is consistent.}$$

I start with considering the conjunctive consistency condition. I take this condition to be plausible because I can not think of a plausible behavior that violates it. Any violation of the condition seems to rather stem from the fact that our initial hypothesis about the belief function of the subject is wrong than that it would show that splitting interpretability imposes too strong constraints on linguistic behaviors.

To discuss the strength of conjunctive covering condition I split it into two separate condition which are easier to understand. This is however only possible under the additional assumption that the set $b[S] = \{b(s) \subseteq W \mid s \in S\}$ is closed under finite intersections and arbitrary unions. This means that for all situations $s_1, \dots, s_n \in S$ there is a situation $c \in S$ such that $b(c) = b(s_1) \cap \dots \cap b(s_n)$, and that for every set of situations $T \subseteq S$ there is a situation $d \in S$ such that $b(d) = \bigcup\{b(t) \mid t \in T\}$. Again it is not plausible that this assumption is always satisfied but I am making it here because it allows for a better understanding of the conjunctive covering condition.

In Proposition 7.2.1 it is verified that if $b[S]$ is closed under finite intersections and arbitrary unions then the conjunctive covering condition is equivalent to the conjunction of the exact covering condition from Definition 2.6.4 and the following new condition:

2.6.8. Definition. A behavior $a : S \to \mathcal{PB}$ satisfies the *conjunction condition* relative to the belief function $b : S \to \mathcal{P}W$ if for all $s, t_1, \dots, t_n \in S$

$$b(s) = b(t_1) \cap \dots \cap b(t_n) \text{ implies } \mathsf{cl}\,(a(t_1) \cup \dots \cup a(t_n)) \subseteq a(s).$$

I already discuss the exact covering condition above for the case where $\mathcal{V} = \mathsf{At}$. Let us consider the conjunction condition.

First, it is shown in Proposition 7.2.1 that the conjunction condition entails the monotonicity condition from Definition 2.6.5.

Second, there is a simple but important instance of conjunction condition where $n = 1$ and $t_1 = s$. In this case the antecedent of conjunction condition is $b(s) \subseteq b(s)$ which is always satisfied. The consequent is that $\mathsf{cl}\,(a(s)) \subseteq a(s)$ which is equivalent to saying that $a(s)$ is a theory. This shows that also the account of this section entails that the set of sentences that the subject accepts in some situation needs to be closed under logical consequence. As before I take this to be a somewhat implausible condition on interpretability that is however difficult to avoid.

The only plausible linguistic behaviors that I can think of that do not satisfy the conjunction condition are cases in which the subject would violate the constraint that the set of sentences that she accepts is closed under logical consequence. In other cases it seems to me that if a linguistic behavior does not satisfy the conjunction condition it is not because the notion of splitting interpretability is too strong but rather because the chosen belief function b does not really capture the beliefs of the subject.

Because I can not think of any convincing counterexamples to any of the condition discussed above I take splitting interpretability with a belief function to satisfy the variety requirement.

Tight interpretability is stronger than splitting interpretability. Hence all the constraints on splitting interpretability discussed here also apply to tight interpretability. I do not have a general characterization of the behaviors that are tightly interpretable with a belief function. Theorem 7.3.10 is an attempt, but it relies on additional assumptions that are quite implausible. I do not discuss this here in any detail. Let me just mention that also in the multi-situation setting tight interpretability with a belief function imposes a cardinality condition that is similar to the constraints for tight interpretability discussed in Section 2.3. Using the notion of an n-theory from Chapter 7 it can be stated as follows:

2.6.9. Definition. A linguistic behavior $a : S \to \mathcal{PB}$ satisfies the *cardinality condition* relative to a belief function $b : S \to \mathcal{P}W$ if for all $s \in S$ it holds that if $b(s)$ is a finite set with n elements then $a(s)$ is an n-theory.

Because it entails the cardinality condition tight interpretability with a belief function is prone to the kind of counterexamples discussed at the end of Section 2.3. The behavior of the subject is not tightly interpretable if there is some situation where she has complete information about the basic facts but is using a vague language. In Section 5.2 I discuss a possibility of improving tight interpretability with a belief function such that it avoids this problem.

To summarize, I take tight interpretability for the case where $\mathcal{V} = \mathsf{At}$ and splitting interpretability for the case where $\mathcal{V} = \mathcal{B}$ to satisfy the variety requirement, because they impose reasonable constraints on linguistic behaviors.

The notions of tight and splitting interpretability from this section satisfy the determinacy requirement. It is clear how to understand the worlds of an interpreting model. In the case of tight interpretability they are just the worlds that we have fixed in advance to specify the belief function of the subject. In the case of splitting interpretability they are related by the splitting function to the worlds that have been given in advance. Hence we do not have the kind of indeterminacy that is observed in Section 2.1 Moreover, the account is designed such that it reduces the indeterminacy for the interpretation function that we observed at the end of Section 2.3. By interpreting the subject across situations in which she has sufficiently distinct belief sets we are able to narrow down the propositions that her beliefs might express according to an interpreting interpretation function.

The major shortcoming of the account of interpretation from this section is the assumption that in every situation, where we are interpreting the subject, we know all of her beliefs about the relevant basic facts. This is too strong if we think of scenarios of radical interpretation. Hence I take the account of this section to not satisfy the little-input requirement. Chapter 3 addresses this problem and explores ways to weaken the assumption that we know all of the beliefs of the subject to the weaker one that we know the perceptual beliefs that the subject has about her environment.

In later chapters of this thesis, I am mainly concerned with different possibilities for modeling of meaning. In most of the discussion there it is unproblematic to assume that we know all of the beliefs that the subject has about the basic facts. For this reason, I use the account of interpretation given in this section as the basis for extension in these later chapters, and not the more complex one developed in Chapter 3.

Chapter 3

From evidence to belief

In this chapter I discuss the problem of interpretation that only presupposes knowledge about the perceptual beliefs of the subject. The setting from this chapter is an adaption of the account from Section 2.6 where it is assumed that we know all the beliefs of the subject. Because in this chapter we only assume to know the perceptual beliefs of the subject, the accounts developed in this chapter perform better on the little-input requirement than the account from Section 2.6.

In Section 3.1 I am introducing the basic idea of this chapter which is to use an evidence function instead of a belief function to encode our assumptions about the beliefs of the subject. In Section 3.2 I then discuss a simple proposal for how the subject forms her beliefs given her evidence about the world. However, as an account of interpretation, this proposal fails to meet the variety requirement. In Section 3.3 I give a more refined account that performs reasonably well on all three requirements for an account of interpretation.

3.1 Evidence functions

In this section I show how to weaken the assumption that we know all of the beliefs of the subject to the weaker one that we know all her perceptual beliefs. To this aim I introduce the notion of an evidence function which represents the perceptual beliefs of the subject across multiple situations.

In many situations it is not plausible that we can determine all the beliefs of the subject without interpreting her linguistic behavior. Let us consider an example. Suppose we are observing the subject looking at the window from inside a building. She sees that there are raindrops on the window but she can not see whether it is raining outside because it is late in the evening and thus dark outside. From just observing the subject in this situation it is difficult to determine whether she believes that it is raining. It might be that she reasons that if there are raindrops on the window then it is likely raining outside. But it could also be that she considers it a relevant possibility that it has just stopped raining and so there are still raindrops on the window but it does not rain. In this case the subject would not believe that it is raining. There seems to be nothing in this situation that tells us how the subject reasons and hence whether she believes that it is raining or she is uncertain to whether it is raining.

There is however something that we do know about the subject in this situation. It is quite plausible to assume that the subject believes that there are raindrops on the window. This is what she can see without any difficulties from the inside of the building.

Instead of assuming that we know all the beliefs of the subject we might more plausibly assume that we only know her perceptual beliefs. In the following I try to give a formal account of interpretation based on this weaker assumption.

The account given here uses the setting from Section 2.5 in which we are interpreting the subject's behavior across multiple situations. We again have the set S of all the situations in which we are interpreting the subject. For every situation we suppose that we know all of the sentences that she accepts in this situation. Hence linguistic behavior is, as in Definition 2.5.1, a function $a : S \to \mathcal{P}\mathcal{V}$ which gives us for every situation $s \in S$ the set $a(s) \subseteq \mathcal{V}$ of all sentences that the subject accepts in the situation s.

We also need a formal representation of our knowledge about the perceptual beliefs that the subject has about her surroundings. Every belief that the subject has in some situation corresponds to a set of possible worlds. Hence we could represent all the perceptual beliefs of the subject by a set of sets of possible worlds. However, there is an a simpler structure that suffices for our purposes.

To determine the subject's linguistic behavior we only interested in the belief set of the subject. In a situation where the subject has certain perceptual beliefs this belief set is a subset of each of these perceptual beliefs. A belief set is a subset of all the elements in a set of sets of worlds exactly if it is a subset of the intersection of all of the elements in the set of sets of worlds. Hence, in order to represent the subject perceptual beliefs it suffices for our purposes to use just a set of possible worlds which we think of as the conjunction of all her perceptual beliefs. This leads to the following definition:

3.1.1. DEFINITION. An *evidence function* is a function $e : S \to \mathcal{P}W$ which maps a situation s to a set of worlds $e(s) \subseteq W$. The set $e(s)$ is the *evidence set* of the subject in the situation $s \in S$.

Intuitively, one can think of $e(s)$ either as the intersection of all the perceptual beliefs of the subject in the situation s or as the set of all possible worlds that are compatible with all the perceptual beliefs that the subject has in s.

My discussion presupposes a notion of evidence and perceptual beliefs on which the subject does not normally have contradictory evidence or contradictory perceptual beliefs. If in a situation s the subject had some perceptual belief $P \subseteq W$ and another perceptual belief Q such that $Q \subseteq W \setminus P$ then it would follow that her evidence set $e(s) \subseteq P \cap Q \subseteq \emptyset$ is empty. Hence our representation of evidence would trivialize in such cases.

Evidence functions are the same type of mathematical object as the belief functions from Definition 2.5.2. The difference is in what we take this functions to represent. If b is the belief function of the subject then the set $b(s) \subseteq W$ is the intersection of all her beliefs in situation s. If e is an evidence function then the set $e(s) \subseteq W$ is the intersection of all the perceptual belief that the subject has in the situation s.

As an example of an evidence function consider again the example from above. There we are only interested in one situation $s \in S = \{s\}$, where the subject sees the raindrops on the window but can not see whether it is actually raining outside because it is dark outside and she is inside a building where the lights are on. To represent her perceptual beliefs in this situation we use the domain $W = \{w, v, u\}$ where it is raining only at w and there are raindrops on the window at w and v but not at u. Because the subject sees the raindrops on the window we know that she has the perceptual belief $\{w, v\}$. And because she does not see whether it is raining we know that she does not have any stronger perceptual belief such as $\{w\}$, which would mean that she sees that it is raining, or $\{v\}$, which would mean that she sees that it is not raining. Hence her strongest perceptual belief is $\{w, v\}$, and so we define her evidence function such that $e(s) = \{w, v\}$.

In this example we are not assuming to know the subject's actual belief set. We only assume that it is a subset of $\{w, v\}$ and hence the subject believes that there are raindrops on the window. It might be that the subject remains undecided to whether it is raining in which case here belief set would be $\{w, v\}$. It could also be that she takes the raindrops on the window to be sufficient reason to belief that it is raining in which case her belief set would be just $\{w\}$. In a unlikely case she might even believe that it is not raining, and so her belief set is $\{v\}$. Maybe she is always optimistic about the weather and believes that it is not raining unless she has direct evidence to the contrary.

In the following two sections of this chapter I am considering different ideas of how one can define interpretability with an evidence function. Both approaches use the multi-situation possible world models from Definition 2.5.3 to interpret the subject's behavior. We need some way of determining the belief function $b : S \to \mathcal{P}W$ of such a model from the evidence function $e : S \to \mathcal{P}W$ that we assume to be given. Obviously, we want that the subject believes all her perceptual beliefs meaning that $b(s) \subseteq e(s)$ for all situations $s \in S$. The approaches of the following two sections differ in how they determine the subset $b(s)$ of $e(s)$.

References to the literature

The assumption that we the interpreters know the perceptual beliefs of the subject seems to correspond to the idea of triangulation in Davidson's later writings (see for instance Davidson 1992). Davidson's idea is that when observing that the subject shows a similar reaction to some stimuli as the interpreter then the subject's perceptual beliefs are about the object that is the common source of the two causal chains that lead to these reactions. In this thesis I am not assuming that the interpreter uses this concrete causal mechanism to determine the subject's perceptual beliefs. I just need that there is some way of determining them.

Lewis (1996, p. 553) discusses a notion of evidence on which the subject can not have contradictory perceptual beliefs. Adapting his terminology somewhat, Lewis defines a world w to be uneliminated in a situation s if and only if the subject's perceptual experience and memory in s at the world w exactly match her perceptual experience and memory in s at the actual world. With this definition the actual world is always uneliminated. Hence, if we define the evidence set of the subject in some situation to be the set of all worlds uneliminated in s then the subject never has contradictory evidence because the actual world is always in the evidence set.

Lewis' notion of evidence fits well to the setting of this chapter. But there are two clarifications that need to be made.

First, Lewis has a much richer conception of possible worlds than the one from this thesis. To say whether the subject's perceptual experience and memory in some world match her perceptual experience and memory in the actual world, we have to know for every world what the subject's perceptual experience and memory are at the world. Hence these facts about the subject's perceptual experience and memory should be part of the basic facts for which we construct the domain of possible worlds. Practically this is not needed because we know intuitively what the subject would normally believe if a world is as described by the basic facts. However, if we want to interpret the subject in some extraordinary situations, as for instance a case where a daemon is messing with her experience, then it might be crucial to include corresponding worlds in the domain.

Second, Lewis claims that if a certain experience eliminates a world then this need not be because the propositional content of the experience conflicts with the world. Hence it remains a possibility for Lewis that the propositional contents of the subject's experiences are contradictory. I am not sure what notion of propositional content of an experience Lewis has in mind here. He does not use this notion at any other place in his paper. But this notion of the propositional content of an experience needs to be distinguished from what I call the perceptual belief that the subject obtains from the experience, because I take the belief set of the subject to be the intersection of all her perceptual beliefs which should usually be non-empty. It is most natural for me to let the perceptual belief obtained by some experience be the set of worlds that are uneliminated by the experience.

3.2 Prior beliefs

In this section I consider an account of interpretation with an evidence function. The idea is that the subject has some prior beliefs independent of her evidence about the particular situation that she is in. Her actual beliefs in some situation are then these prior beliefs together with all the perceptual beliefs that she has about the situation. The resulting account of interpretation is simple but faces serious difficulties with the variety requirement.

To illustrate the idea of prior beliefs consider again the example from the previous section. There we imagine a situation where the subject can see the raindrops on the window but, since she is inside a building and it is dark outside, she can not see whether it is actually raining outside. In this situation her perceptual beliefs do not determine whether it is raining. The subject might however still believe that it is raining because she believes, independently of her evidence in this situation, that if there are raindrops on the window then it is raining.

I call such beliefs that the subject has independently of her evidence about her situation the *prior beliefs* of the subject. In the example above the belief that if there are raindrops on the window then it is raining is a prior belief of the subject. We can model the prior beliefs of the subject with a set of worlds $U \subseteq W$. This set U is the conjunction of all of the prior beliefs of the subject, or equivalently, it contains all the worlds in which all the prior beliefs are true.

The subject's actual beliefs in some situation are all propositions that are implied by the conjunction of her prior beliefs and her perceptual beliefs about the situation. Therefore the belief set that the subject has in some situation is the set of all worlds that are compatible with her evidence about the situation and that make all of her prior beliefs true. Formally this means that given a set $U \subseteq W$ representing the subject's prior beliefs and an evidence function $e : S \to \mathcal{P}W$ we define the belief function $b : S \to \mathcal{P}W$ of the subject such that $b(s) = U \cap e(s)$ for all situations $s \in S$.

Think again of the situation where the subject sees the raindrops on the window but she can not determine whether it is raining. To model her beliefs in this situation, which we call s_D, take the domain $W = \{w, v, u\}$ such that w is the only world where it is raining and at w and v but not at u there are raindrops on the window. The perceptual beliefs of the subject are that there are raindrops on the window and nothing stronger. Hence $e(s_D) = \{w, v\}$. Her only prior belief is that if there are raindrops on the window then it is raining. Hence $U = \{w, u\}$. Combining her prior beliefs with her perceptual beliefs yields that the subject believes that it is raining because $b(s_D) = \{w, u\} \cap \{w, v\} = \{w\}$.

What distinguishes the account of this section from the one in Section 2.6 is that we do not assume to know the set U in advance to interpretation. We only know the evidence function e and try to infer the prior beliefs U, and consequently also the belief function b, from the subject's linguistic behavior. Hence when interpreting the subject we try to deduce her prior belief set $U \subseteq W$ and the interpretation function $I : \mathsf{At} \to \mathcal{P}W$ that encodes meanings in the language of the subject. This leads to the following definition:

3.2.1. Definition. A *prior belief model* is a triple (W, U, I), where W is a domain of worlds, $U \subseteq W$ is the prior belief set and $I : S \to \mathcal{P}W$ is the interpretation function.

Using the definition of a belief function from a prior belief set and an evidence function we can now define a notion of tight interpretability:

3.2.2. Definition. A prior belief model (W, U, I) *interprets a linguistic behavior* $a : S \to \mathcal{PV}$ *with an evidence function* $e : S \to \mathcal{P}W$ if $a = a^M$ for the multi-situation model $M = (W, b, I)$, where $b : S \to \mathcal{P}W$ is defined such that $b(s) = U \cap e(s)$ for all $s \in S$.

A behavior $a : S \to \mathcal{PV}$ is *tightly prior belief interpretable with an evidence function* $e : S \to \mathcal{P}W$ if there is some prior belief model that interprets a with e.

We can also define a notion of splitting interpretability that is better behaved than tight prior belief interpretability. For this we need to add a splitting function to the definition of the interpreting models:

3.2.3. Definition. A *splitting prior belief model* over W is a tuple (W', f, U, I), such that (W', U, I) is a prior belief model and (W', f, I) is a splitting interpretation function over W.

For defining splitting prior belief interpretability we have the problem that the belief function $e : S \to \mathcal{P}W$ is defined relative to the original domain whereas the interpreting splitting prior belief model (W', f, U, I) over W has the splitting W' as its domain. Similarly as in Section 2.4 we can solve this problem by using the lifting $f^{-1}[e(s)] \subseteq W'$ of the evidence set $e(s) \subseteq W$ in some situation $s \in S$.

3.2.4. Definition. A splitting prior belief model (W', f, U, I) *interprets a linguistic behavior* $a : S \to \mathcal{PV}$ *with an evidence function* $e : S \to \mathcal{P}W$ if $a = a^M$ for the multi-situation possible world model $M = (W', b', I)$, where $b' : S \to \mathcal{P}W'$ is defined such that $b'(s) = U \cap f^{-1}[e(s)]$ for all $s \in S$.

A behavior $a : S \to \mathcal{PV}$ is *splitting prior belief interpretable with an evidence function* $e : S \to \mathcal{P}W$ if there is some splitting prior belief model that interprets a with e.

Let us consider an example which shows how the prior beliefs of the subject can be determined by her linguistic behavior. As before we use the domain $W = \{w, v, u\}$ with three worlds where it is raining only at w and there are raindrops on the window at w and v. We observe the subject's behavior in two situations. In s_D she is inside a building, seeing the raindrops on the window but unable to see whether it is raining outside. Hence her evidence function is such that $e(s_D) = \{w, v\}$. Assume that in this situation she is accepting the sentence p and all its logical consequences. Hence her linguistic behavior is such that $a(s_D) = \mathsf{cl}\,(\{p\})$. In the other situation s_B the subject is sunbathing on some beach. She clearly knows that it is not raining but she has no evidence to whether there are raindrops on the windows of the building inside which she was in situation s_D. Hence we have that $e(s_B) = \{v, u\}$. Also assume that in this situation the subject accepts the logical consequences of $\neg p$, so $a(s_B) = \mathsf{cl}\,(\{\neg p\})$.

The linguistic behavior from this example is tightly prior belief interpretable with the given evidence function. This is witnessed by the prior belief model (W, U, I), where $U = \{w, u\}$ and $I : \mathsf{At} \to \mathcal{P}W$ is defined such that $I(p) = \{w\}$. This shows that one way to interpret the subject in this example is to assume that she has the prior belief that if there are raindrops on the window then it is raining.

There is a sense in which the information available in the above example determines that the subject has the prior belief that if there are raindrops on the window then it is raining. Whenever a splitting prior belief model (W', f, U, I) interprets the behavior in the example with the evidence function in the example then $U \subseteq f^{-1}[\{w, u\}]$. To see this assume for a contradiction that there is some $v' \in U$ such that $f(v') = v$, hence there is a world v' compatible with the subject's prior beliefs in which there are raindrops on the window but it is not raining. Then it would follow that $v' \in U \cap f^{-1}[e(s_D)] = b(s_D)$ and that $v' \in U \cap f^{-1}[e(s_B)] = b(s_B)$. This means that v' is a doxastic alternative for the subject in the situation where she is inside the building and just sees the raindrops and it is a doxastic alternative in the situation where she is sunbathing on the beach. But since the subject accepts p in s_D we have that $b'(s_D) \subseteq I(p)$ and hence $v' \in I(p)$. Similarly because she accepts $\neg p$ in s_B it follows that $b'(s_B) \subseteq I(\neg p)$ and hence that $v' \in I(\neg p)$. But this is a contradiction since p can not be both true and false at v'.

The above example suggest that the account given in this section fares reasonably well with respect to the determinacy requirement. The linguistic behavior in the example determines the prior beliefs of the subject and the interpretation function at the worlds that are compatible with these prior beliefs.

Also with respect to the little-input requirement the account is doing quite well. We only assume to know the sentences that the subject accepts and the beliefs that she obtains from perception. This weakens the assumption from Section 2.6 that we know all of the subject beliefs.

Prior belief interpretability with an evidence function does not satisfy the variety requirement. There are certain behaviors that a subject might quite plausibly show that are not prior belief interpretable.

As an example consider a variation on the example above. We are interpreting the subject in two situation s_D and s_W. As above s_D is a situation where the subject sees from inside a building that there are raindrops on the window. Since it is dark outside and the lights are on she can not see whether it is actually raining outside. Relative to our domain $\{w, v, u\}$, where w is the only rainy world and w and v are the worlds where there are raindrops on the window, this means that the evidence of the subject is such that $e(s_D) = \{w, v\}$. Also assume that in this situation the subject actually accepts the sentence p and all of its logical consequences, that is, $a(s_D) = \mathsf{cl}(\{p\})$. Now let us suppose that the subject is actually curious to whether it is raining outside, maybe because she is planning to leave the building. So she gets closer to the window to be able to see whether it is raining outside. She notices that it has just stopped raining, even though there are still the raindrops on the window. In this second situation s_W, where the subject is standing at the window and looking outside, her evidence is $e(s_W) = \{v\}$. Also suppose that in this situation the subject no longer accepts p and even comes to accept the negation of p and all its consequences. Formally this is encoded by setting $a(s_W) = \mathsf{cl}(\{\neg p\})$.

It is quite plausible that a subject shows this behavior. Think of the sentence p as meaning that it is raining. When in s_D the subject is just looking at the window and seeing the raindrops she infers that it is raining from the fact that there are raindrops on the window. Hence she accepts the sentence p in the situation s_D. Later when she is going closer to the window she notices that it is actually not raining. Now she believes that it is not raining and hence accepts $\neg p$. She also does not accepts p anymore because she no longer believes that it is raining.

The behavior a from the example above is not splitting interpretable with the evidence function e. Intuitively, the reason is that in s_D the subject only comes to believe that it is raining if she has the prior belief that if there are raindrops on the window then it is raining. However, this prior belief is incompatible with the evidence that subject has in s_W, which is that there are raindrops on the window but it is not raining. Hence her belief set in s_W would be empty and as a consequence she would accepts all sentences. This does not happen to reasonable subject.

To prove that the account from this section can not cope with the kind of example given here we can use the representation results for prior belief interpretability.

For the case $\mathcal{V} = \mathsf{At}$, where the linguistic behavior just contains atomic sentences, it is shown in Theorem 7.3.2 that tight and splitting prior belief interpretability with an evidence function $e : S \to \mathcal{P}W$ are the same as tight and splitting interpretability with e in the sense of Section 2.6, where e used as if it were a belief function. This also means that a behavior is prior belief interpretable with an evidence function if and only if it satisfies the simple covering condition from Definition 2.6.3 relative to the evidence function.

For case $\mathcal{V} = \mathcal{B}$, so the subject is accepting propositional formulas, Theorem 7.3.4 shows that a behavior is splitting prior belief interpretable with an evidence function if and only if it satisfies the conjunctive covering condition from Definition 2.6.6. The conjunctive covering condition is one of the conditions for splitting interpretability discussed in Section 2.6. Splitting prior belief interpretability is similar to splitting interpretability from Section 2.6. The only difference is that splitting prior belief interpretability does not require the conjunctive consistency condition. Similarly, it is shown in Theorem 7.3.10 that under additional assumptions tight prior belief interpretability is the same as tight interpretability minus the conjunctive consistency condition. Hence also tight prior belief interpretability implies the conjunctive covering condition.

The simple covering condition and the conjunctive covering condition are relevant for the discussion here because, as observed in Section 2.6, they imply the monotonicity condition from Definition 2.6.5. Relative to an evidence function $e : S \to \mathcal{P}W$ the monotonicity condition requires that for all $s, t \in S$

$$e(s) \subseteq e(t) \text{ implies } a(t) \subseteq a(s).$$

This condition says that if in some situation s the subject has more evidence about the world than in some other situation t then every sentence that she already accepts in t she also accepts in s.

The example above shows that it is not plausible that every linguistic behavior satisfies the monotonicity condition. In the situation s_W, where the subject is standing at the window and seeing that it is not raining outside, she has stronger perceptual beliefs than in the situation s_D where she merely sees the raindrops on the window. However, she accepts the sentence p in the situation s_D and no longer accepts it in s_W. Hence her behavior does not satisfy the monotonicity condition and therefore the account of this section does not satisfy the variety requirement.

In the next section I discuss another approach for relating the belief set of the subject to her evidence set that does not validate the monotonicity condition.

3.3 Plausibility orders

In this section I consider an approach for relating the beliefs of the subject to the evidence that is less rigid than the setting from the previous section and hence does not have the same difficulties with the variety requirement. The resulting account of interpretation performs reasonably well on three requirements from Section 1.4. It is probably the most balanced account of interpretation that I discuss in this thesis.

The idea of this section is that the subject has a prior plausibility ordering $\leq \subseteq W \times W$ over the set of all possible worlds W. The intuition is that whenever $w \leq v$ holds for two worlds w and v then in every situation where the subject has evidence that is compatible with both w and v she considers it at least as plausible that w rather than v is the actual world.

The doxastic alternatives of the subject in some situation are then all the worlds that are compatible with her evidence in the situation and that are most plausible with respect to her prior plausibility ordering. Formally this means that $b(s) = \mathsf{Min}_{\leq}(e(s))$ for all situations $s \in S$, where $\mathsf{Min}_{\leq}(X)$ for some $X \subseteq W$ is defined such that

$$\mathsf{Min}_{\leq}(X) = \{m \in W \mid m \leq w \text{ for all } w \in X \text{ with } w \leq m\}.$$

For this notion of minimal elements to work well one needs to require some properties of the plausibility ordering $\leq$. At least $\leq$ should be reflexive and transitive and well-founded in the sense that there is no infinite chain $w_1 \geq w_2 \geq \dots$ such that $w_1 \not\leq w_2 \not\leq \dots$. Such reflexive, transitive and well-founded relations are called well-founded preorders. One might also require that $\leq$ is antisymmetric, meaning that it is a well-founded poset. Whether one considers all well-founded preorders or just well-founded posets has no influence on the representation theorem for the account of interpretation given in this section.

Let us have a look at an example of the kind discussed at the end of Section 3.2. We consider three worlds $W = \{w, v, u\}$ such that at u it is raining and at u and v there are raindrops on the window. Suppose that the subject has the plausibility order $\leq = \{(w, w), (v, v), (u, u), (w, u)\}$. With this plausibility order the subject considers the world w more plausible than u. She thinks that if there are raindrops on the window then it is more plausible that it is raining than that it is not. In the situation s_D, where she sees the raindrops on the window but can not see whether it is raining outside, she still believes that it is raining outside because $b(s_D) = \mathsf{Min}_{\leq}(e(s_D)) = \mathsf{Min}_{\leq}(\{w, u\}) = \{w\}$. But when she now goes to the window and sees that it just stopped raining then her beliefs do not become inconsistent because $b(s_W) = \mathsf{Min}_{\leq}(e(s_W)) = \mathsf{Min}_{\leq}(\{u\}) = \{u\}$.

The account of interpretation should determine the prior plausibility order of the subject from her linguistic behavior. Hence an interpreting model needs to contain a plausibility order and an interpretation function. If one defines an interpreting model to be just a pair consisting of a plausibility order and an interpretation function then one would obtain an account of interpretation for plausibility orders that corresponds to tight interpretability as discussed in Section 2.5. I have however not been able to obtain a good characterization of tight interpretability for plausibility orders, and so I only discuss splitting interpretability. For splitting interpretability we need to add a splitting function, as discussed in Section 2.4 to the definition of an interpreting model:

3.3.1. Definition. A *splitting plausibility model over* W is a tuple $(W', f, \leq, I)$ such that, W' is the domain of worlds, $f : W' \to W$ is a function that need not be surjective, $\leq \subseteq W' \times W'$ is a well-founded preorder on W' and $I : \mathsf{At} \to \mathcal{P}W'$ an interpretation function.

Note that because in the above definition $f : W' \to W$ is not required to be surjective it is in general not the case that if $(W', f, \leq, I)$ is a splitting plausibility model over W then (W', f, I) is a splitting interpretation model over W in the sense of Definition 2.4.2. That we allow f to not be surjective means intuitively that we allow the subject to forget about some of the possible worlds in the original domain. This is similar to prior belief interpretability as discussed in the previous section. My reason to not assume that f is surjective is to simplify the conditions for plausibility interpretability, which are already difficult enough. If we were to require that the function $f : W' \to W$ is surjective then we would need to add a consistency condition, similar to the conjunctive consistency condition, to the representation theorem discussed below.

When defining the notion of splitting interpretability with an evidence function we again run into the difficulty that the evidence function $e : S \to \mathcal{P}W$ is defined relative to a different set of worlds than the interpretation function in an interpreting splitting plausibility model. As discussed in Section 2.4, the solution is to work with the lifting $f^{-1}[e(s)] \subseteq W'$ of the evidence set $e(s) \subseteq W$. We then obtain the following definition of interpretability:

3.3.2. Definition. A splitting plausibility model $(W', f, \leq, I)$ *interprets a linguistic behavior* $a : S \to \mathcal{PV}$ *with an evidence function* $e : S \to \mathcal{P}W$ if $a = a^M$ for the multi-situation possible world model $M = (W', b', I)$, where $b' : S \to \mathcal{P}W'$ is defined such that $b'(s) = \mathsf{Min}_{\leq}(f^{-1}[e(s)])$ for all $s \in S$.

A behavior $a : S \to \mathcal{PV}$ is *splitting plausibility interpretable with an evidence function* $e : S \to \mathcal{P}W$ if there is some splitting plausibility model that interprets a with e.

Since I do not discuss tight plausibility interpretability I usually call splitting plausibility interpretability just plausibility interpretability.

As an example of plausibility interpretability let us reconsider the example form Section 3.2 of a behavior that does not satisfy the monotonicity condition and hence is not interpretable in the account of that section. We again use the domain $W = \{w, v, u\}$, where it is raining only at w and there are raindrop on the window at w and v but not at u. We observe the subject in two situation s_D and s_W. In s_D the subject is looking at the window from inside a building and because it is dark outside she can not see whether it is raining. She however sees that there are raindrops on the window. The subject perceptual beliefs in this situation are encoded by $e(s_D) = \{w, v\}$. We assume that in this situation the subject accepts the sentence p and all of its consequences, that is, $a(s_D) = \mathsf{cl}(\{p\})$. In situation s_W the subject went to the window to look more carefully whether it is raining outside. She sees that it has just stopped raining so that there are still raindrops on the window. Hence we have that $e(s_W) = \{v\}$. With this new additional evidence the subject no longer accepts the sentence p and in fact comes to accept all the consequence of the sentence $\neg p$. Hence we have $a(s_W) = \mathsf{cl}(\{\neg p\})$.

This example is not interpretable with the account of Section 3.2. It does not satisfy the monotonicity condition because in s_W the subject has strictly more information than in s_D, that is, $e(s_W) \subseteq e(s_D)$, but in s_W the subject no longer accepts p which she did accept in s_D. On the account of this section the behavior in the example is interpretable. An interpreting model can be based on the plausibility order given in the example above. Consider the splitting plausibility model $M = (W, f, \leq, I)$, where $f : W \to W$ is the identity function, $\leq = \{(w, w), (v, v), (u, u), (w, u)\}$ and $I(p) = \{w\}$, which is the meaning that the sentence "It is raining." has in English. To see that this model indeed interprets the behavior a of the subject observe that $b(s_D) = \mathsf{Min}_{\leq}(\{w, u\}) = \{w\} \subseteq I(p)$ and $b(s_W) = \mathsf{Min}_{\leq}(\{u\}) = \{u\} \subseteq I(\neg p)$.

Let us now see how plausibility interpretability performs on the three requirements from Section 1.4.

For the variety requirement the situation is the same as for prior belief interpretability from the previous section. We are only assuming that we know the perceptual beliefs of the subject in every situation, which is quite modest.

The account also satisfies the determinacy requirement. At least it seems to me that in all examples where the linguistic behavior is rich enough the interpreting plausibility order is reasonably determinate.

I already show above that the counterexample against prior belief interpretability is plausibility interpretable. I can not think of any other example of a linguistic behavior that is not plausibility interpretable. Hence it seems that the account also satisfies the variety requirement.

In the remainder of this section I explain the representation results that characterizes the class of plausibility interpretable behaviors.

First consider the case where $\mathcal{V} = \mathsf{At}$, hence we are not making any assumptions about what expressions function as the propositional connectives in the language of the subject. For this case Theorem 7.4.4 shows that the plausibility interpretable behaviors are precisely those that satisfy the exact covering condition from Definition 2.6.4.

The characterization of the plausibility interpretable behaviors is more involved in the case where $\mathcal{V} = \mathcal{B}$ and so we have a hypothesis about what expressions are the propositional connectives in the language of the subject. To state the condition for plausibility interpretability of a behavior $a : S \to \mathcal{PB}$ with an evidence function $e : S \to \mathcal{P}W$ we need the following auxiliary definitions:

3.3.3. Definition. Fix a linguistic behavior $a : S \to \mathcal{PB}$ and an evidence function $e : S \to \mathcal{P}W$. For some given formula $\varphi \in \mathcal{B}$ we say that a world $w \in W$ *is potentially φ in a set of worlds $X \subseteq W$* if there are $s_1, \dots, s_n \in S$ such that $w \in e(s_i)$ and $e(s_i) \subseteq X$ for all $i \in \{1, \dots, n\}$ and $\varphi \in \mathsf{cl}\,(a(s_1) \cup \dots \cup a(s_n))$.

A world $w \in W$ *is implausible in a set of worlds $X \subseteq W$* if w is potentially $\bot$ in X.

That a world w is potentially φ in a set of worlds X should be understood such that, according to the linguistic behavior of the subject, if the prior plausibility order of the subject is such that w is a minimal element of X then the proposition expressed by φ in the language of the subject needs to be true at w. The reason is that if w is minimal in X then it also needs to be minimal for any of its subsets

$e(s_i) \subseteq X$. Therefore all of the sentences that the subject accepts in any such s_i are true at w and hence also any logical consequence of these sentences.

If a world w is implausible in a set of worlds X it follows that, according to the linguistic behavior of the subject, w can not be a minimal element of X in the prior plausibility order of the subject. Because if w was minimal in X then by the reasoning of the previous paragraph it would have to be the case that the proposition expressed by $\perp$ is true at w, which is impossible.

An example of the latter notion can be found in the linguistic behavior discussed above, where $a(s_D) = \mathsf{cl}(\{p\})$, $a(s_W) = \mathsf{cl}(\{\neg p\})$, $e(s_D) = \{w, u\}$ and $e(s_w) = u$. Here the world u, at which there are raindrops on the window but it is not raining, is implausible in the set $\{w, u\}$ of the worlds at which it is raining. This is the case because $w \in e(s_D) \subseteq \{w, u\}$, $w \in e(s_W) \subseteq \{w, u\}$, $p \in a(s_D)$, $\neg p \in a(s_w)$ and $p, \neg p \models \perp$.

We can now state the representation result. Theorem 7.4.5 shows that a behavior $a : S \to \mathcal{PB}$ is plausibility interpretable with an evidence function $e : S \to \mathcal{P}W$ if and only if it satisfies the following plausibility covering condition:

3.3.4. Definition. A behavior $a : S \to \mathcal{PB}$ satisfies the *plausibility covering condition* relative to an evidence function $e : S \to \mathcal{P}W$ if for all $\varphi \in \mathcal{B}$ and $s \in S$ it holds that $\varphi \in a(s)$ whenever there is an $X \subseteq W$ such that $e(s) \subseteq X$, all $w \in X \setminus e(s)$ are implausible in X and all $w \in e(s)$ are potentially φ in X.

First note that this condition implies that for every $s \in S$ the set of sentences $a(s)$ that the subject accepts in s is a theory. To see that this is the case assume that we have $\Sigma \models \varphi$ and that $\Sigma \subseteq a(s)$. We want to show that then $\varphi \in a(s)$. Clearly we have that $\varphi \in \mathsf{cl}(a(s))$ because $\Sigma \models \varphi$ and $\Sigma \subseteq a(s)$. By Definition 3.3.3 it follows that φ is plausible at every $w \in e(s)$. Hence $\varphi \in e(s)$ follows from the plausibility covering condition in the case where $X = e(s)$.

The strength the plausibility covering condition is in between the conjunctive covering condition and the exact covering condition. For $\varphi \in a(s)$ to hold one needs that every world in $a(s)$ can be covered with evidence sets that belong to situation in which the subject accepts formulas that have φ as a logical consequence. This is similar to the conjunctive covering condition. But more like the exact covering condition these covering evidence sets almost need to be subsets of $e(s)$, up-to worlds that the subject herself consider to be implausible.

References to the literature

Plausibility orders are a standard semantics for theories of belief revision (see for instance Grove 1988) and for doxastic logics (see for instance Baltag and Smets 2006). The idea of considering the minimal elements in some ordering over possible worlds has originally been proposed by Lewis (1973) as a formal semantics for counterfactual conditionals. A similar approach is also used in artificial intelligence to account for non-monotonic consequence relations (see for instance Kraus, Lehmann, and Magidor 1990).

The difference between the setting of this section and the approaches in belief revision theory and doxastic logic is that in the account given here the language of the subject is determined by interpretation, whereas in belief revision and doxastic logic it is assumed either that the subject speaks the same language as the modeler or that her behavior can be modeled on the level of propositions rather than sentences. From the perspective of this thesis on might think of existing axiomatizations in belief revision theory and doxastic logic as representation results for plausibility models in a setting similar to the one from Section 2.2, where it is assumed that we know the language of the subject prior to interpretation.

In the literature on belief revision and conditional logic mentioned above it is often assumed that the plausibility order $\leq$ is complete, that is, for all worlds w and v at least one of $w \leq v$ and $v \leq w$ holds. In both settings this assumption can be given up, see for instance (Rott 2014) for belief revision and (Veltman 1985) for conditional logic. In this section I do not require the plausibility orders to be complete because I could not find any succinct conditions that characterize splitting plausibility interpretability for this more restricted class of models.

At this point I also want to connect to the principle of charity, which plays a central role in Davidson's work on radical interpretation. In early writings Davidson (for instance Davidson 1973, p. 136) formulates this principle as requiring that the interpreter chooses an interpretation that maximizes the agreement between the subject and him. Davidson suggests that this version of the principle of charity helps to reduce the indeterminacy of interpretation. In the account of interpretation from this section there seems to be no counterpart for this formulation of the principle of charity.

In later work, Davidson (1992) explicitly states that charity is a condition for interpretability and he splits the principle into two parts. The first is the principle of coherence that requires that the subject has a minimal degree of logical consistency in her thoughts. The second is the principle of correspondence and requires that the subject is responding to the same features in her environment as the interpreter would if he was in similar circumstances as the subject. Both these principles seem to have counterparts in the account of this section. On a loose reading of "logic" and "thought" on might take the principle coherence to correspond to the plausibility covering condition. The counterpart of the principle of correspondence might be our assumption that we can know the perceptual beliefs of the subject prior to interpretation.

Chapter 4

Acceptance

In Section 1.3 I introduce the acceptance principle which plays a crucial role in the accounts of interpretation from this thesis. The acceptance principle says that the subject accepts a sentence if and only if she believes the proposition expressed by the sentence. In this chapter I discuss a distinction between two versions of the acceptance principle that result from different ways of making precise what the proposition expressed by some sentence is.

In Section 4.2 I introduce the first version of the acceptance principle, which I call the disquotational acceptance principle. It takes the proposition expressed by a sentence to be determined by the semantic facts which obtain at the actual world.

In Section 4.3 I introduce the second version of the acceptance principle, which I call the metasemantic acceptance principle. It takes the proposition expressed by a sentence to be determined by the subject's beliefs about the semantic facts.

The two different versions of the acceptance principle arise from different ways of relating the proposition expressed by a sentence to the semantic facts which hold at the possible worlds of the domain of some model. To precisely characterize these two acceptance principles I first introduce a class of possible world models in which the semantic facts that obtain at the worlds of the model are explicitly represented. This is the content of Section 4.1

In Section 4.4 I discuss on what grounds we might prefer one version of the acceptance principle over the other. I suggest that the distinction is purely theoretical and that ultimately the choice between disquotational and metasemantic acceptance depends on what mathematical models we find more convenient to work with.

4.1 Semantic facts

In this section I show how semantic facts can be modeled inside the possible world framework. This provides a formal setting in which I can explain the distinction between disquotational and metasemantic acceptance that is the topic of later sections of this chapter.

I assume that it is either true or false that a sentence has a certain meaning. For instance it is a fact that "It is raining." means that it is raining but it is not the case that "It is raining." means that snow is white.

Here one might object that the semantic facts should be relativized to languages. For instance it is only a fact that "It is raining." means in English that it is raining. In most languages this sentence does not have a determinate meaning and we could also imagine a language in which it means something completely different. I do not bother about this relativization to languages because I am only interested in one language, which is whatever language the subject that we are interpreting is speaking. I however shortly return to this issue in an example at the end of this Chapter.

I am calling facts that are about the meaning of sentences *semantic facts*. By calling them so, I do not intend to commit to anything stronger than the claim that it can be true or false that a sentence has a particular meaning. I want the terminology to stay neutral with respect to any particular view about the nature of semantic facts.

When setting up a domain of possible worlds, as described in Section 1.1, one can treat every semantic fact as a kind of basic fact that obtains at some of the worlds in the domain. Consider an example where we have three basic facts, the fact that it is raining, the semantic fact that the sentence p means that it is raining and the semantic fact that p means that it is not raining. Take a domain $W = \{w, v, u, z\}$ with four possible worlds. We specify that it is raining at w and u but not at v and z. Moreover at w and v the sentence p means that it is raining whereas at u and z it means that it is not raining.

Relative to this domain W we could postulate that w is the actual world. This amounts to saying that it is raining and that the sentence p means that it is raining. We could also imagine a subject that has the belief set $B = \{w, v\}$. This subject believes that p means that it is raining but she is uncertain to whether it is raining or not. The subject can also be uncertain about the semantic facts. For instance when she has the belief set $B = \{w, u\}$. In this case the subject believes that it is raining but she is uncertain about the meaning of p. She believes that its meaning is related to the weather but she does not know whether it means that it is raining or that it is not raining.

We can also make the semantic facts explicit in the formal structure of our models instead of specifying them from the outside as we do when treating them like the basic facts from Section 1.1. To formally represent the information given by the semantic facts we might use an interpretation function $I : \mathsf{At} \to \mathcal{P}W$ which maps atomic sentences to the propositions that they express. Because different semantic facts might obtain at different worlds we need an interpretation function for every possible world to represent the semantic facts that hold at that world. Formally, this means that we are using a family $(I_w)_{w \in W}$ of interpretation functions such that for every world w in the domain the interpretation $I_w : \mathsf{At} \to \mathcal{P}W$ specifies the semantic facts that obtain at w.

Let us consider again the example with the domain $W = \{w, v, u, z\}$, where it is raining at w and u but not at v and z. Instead of describing the semantic facts holding at these worlds as above we can now use the family of interpretation functions (I_w, I_v, I_u, I_z) to specify the semantic facts holding at these worlds. We define that $I_w(p) = I_v(p) = \{w, u\}$ and that $I_u(p) = I_z(p) = \{v, z\}$. According to this family of interpretations the semantic facts are such that at w and at v the sentence p expresses the proposition that it is raining, and at u and at z it expresses the proposition that it is not raining.

By using an interpretation function to specify the semantic facts holding at possible worlds we make the assumption that every sentence can only express exactly one proposition. This is not required when we specify the semantic facts as part of the basic facts holding at the possible worlds. I am accepting this assumption for now, but in Section 5.2 I discuss a possibility for weakening it again.

A convenient notation for a family $(I_w)_{w \in W}$ of interpretation functions are tables, in which a row indexed by a world w corresponds to the interpretation I_w such that the cell corresponding to the column v tells us whether $v \in I_w(p)$ for some specified atomic sentence p. Consider again the example above where $W = \{w, v, u, z\}$ and the interpretations are such that $I_w(p) = I_v(p) = \{w, u\}$ and $I_u(p) = I_z(p) = \{v, z\}$. The information given by these interpretation functions corresponds to the following table:

	w	v	u	z
w	p	$\neg p$	p	$\neg p$
v	p	$\neg p$	p	$\neg p$
u	$\neg p$	p	$\neg p$	p
z	$\neg p$	p	$\neg p$	p

We can include the explicit representation of semantic facts by a family of interpretation functions into the definition of a multi-situation model. To do so we replace the single interpretation in a multi-situation possible world model with a family of interpretation functions. This yields the following definition:

4.1.1. Definition. A *metasemantic model* is a tuple $(W, b, (I_w)_{w \in W})$ such that W is a domain of possible worlds, $b : S \to \mathcal{P}W$ a belief function relative to W and $I_w : \mathsf{At} \to \mathcal{P}W$ is an interpretation function for every world $w \in W$ in the domain.

Semantic facts are formally represented by the family of interpretation functions in a metasemantic model. The purpose of the remaining sections of this chapter is to explain how these semantic facts relate to the linguistic behavior of the subject. To this aim I present two alternative definitions of the linguistic behavior generated by some metasemantic model.

References to the literature

Representing the semantic facts that hold at the worlds of a model by using family of interpretation functions in which interpretation functions are indexed by possible worlds yields a framework that is, as far as I can tell, the same as Stalnaker's (2001; 2004) metasemantic interpretation of two-dimensional semantics.

Two-dimensional semantics has been introduced by Kamp (1971) as a formal semantics of the indexical "now" in temporal modal logic. Later, Kaplan (1989) generalized the framework and popularized it in philosophy. In these applications of the two-dimensional framework the meaning of expressions is relativized to contexts which provide the referents for indexical expressions. In this thesis I am not concerned with the problem of accounting for the indexicality or context-dependence of meaning. I am using the two-dimensional framework solely to explicitly represent semantic facts in the setting of epistemic modal logic.

Van Fraassen (1977; 1979) is probably the first to use a two-dimensional semantics with an interpretation that is similar to the one of this section. Stalnaker explicitly develops the metasemantic interpretation in (2001) and (2004) to clarify his preferred reading of his own earlier paper (Stalnaker 1978). In Chapter 6 I compare Stalnaker's understanding of the two-dimensional framework with an alternative that is presented for instance by Chalmers (2002; 2006).

4.2 Disquotational acceptance

In this section I discuss the disquotational acceptance principle. On this version of the acceptance principle the proposition expressed by a sentence in the language of the subject is determined by the semantic facts at the actual world. Thus one obtains a direct connection between the subject's acceptance of sentences and the semantic facts that obtain at the actual world.

The general version of the acceptance principle formulated in Section 1.3 states that the subject accepts some sentence if and only if she believes the proposition expressed by that sentences. This formulation leaves it open what the proposition expressed by some sentence is. The most natural definition in terms of the approach to semantic facts from Section 4.1 is to let the proposition expressed by a sentence be the proposition that the sentence expresses according to the semantic facts holding at the actual world. Thus we obtain the following *disquotational acceptance principle*:

> The subject accepts a sentence if and only if she believes the proposition that the sentence expresses according to the semantic facts that obtain at the actual world.

The formal counterpart of the acceptance principle from Section 1.3 is Definition 2.5.4 of the linguistic behavior generated by some multi-situation model.

Using metasemantic possible world models we can adapt this definition to make precise how the disquotational acceptance principle links the linguistic behavior of the subject to the semantic facts holding at the actual world. To do so we need to know which world in the domain is the actual world. As explained Section 2.5 the actual world depends on the situation in which we are interpreting the subject. Taking this into consideration we obtain the following definition of the behavior disquotationally generated by a metasemantic model:

4.2.1. Definition. The *behavior* $a^M : S \to \mathcal{P}\mathcal{V}$ *disquotationally generated* by a metasemantic model $M = (W, b, (I_w)_{w \in W})$ is defined such that for all situations $s \in S$

$$a^M(s) = \{\varphi \in \mathcal{V} \mid b(s) \subseteq I_{w_s}(\varphi)\},$$

where w_s is the actual world of the situation s.

As an example let us consider again the model $M = (W, b, (I_w)_{w \in W})$ from Section 4.1. The domain $W = \{w, v, u, z\}$ of this model contains four possible worlds such that it is raining at w and u and not raining at v and z. The semantic facts are such that at w and at v the sentence p means that it is raining whereas at u and z it means that it is not raining. This is encoded by the family of interpretation functions $(I_w)_{w \in W}$ such that $I_w(p) = I_v(p) = \{w, u\}$ and $I_u(p) = I_z(p) = \{v, z\}$.

We are just considering one situation s such that in s the subject has the belief set $b(s) = \{w, u\}$. So the subject believes that it is raining but she is uncertain whether p means that it is raining or whether it means that it is not raining. Also assume that w is the actual world w_s of the situation s, hence it is raining and p means that it is raining.

The behavior a^M disquotationally generated by M is such that $p \in a^M(s)$. The reason is that at the actual world the sentence p means that it is raining, which is a proposition that the subject believes in the situation s. That the subject does not know that p means that it is raining is not relevant for her linguistic behavior.

According to the disquotational acceptance principle the semantic facts at the actual world determine the linguistic behavior of the subject. The subject's beliefs about these semantic facts have no influence on her linguistic behavior. This might seem strange. Why would the subject as in the example above accept the sentence p in a situation where she believes that it is raining but considers it possible that p means that it is not raining? One might think that this example shows that there is something wrong with disquotational acceptance.

If one accepts disquotational acceptance then the right reaction to the example above is to think that there is something wrong with the example. On disquotational acceptance the semantic facts of the actual world determine her linguistic behavior. Usually we would expect the subject to know the facts that determine her linguistic behavior. The problem in the above example is that the subject fails to know the semantic facts that determine her own behavior.

At this point it is helpful to recall the analogy between the theory of interpretation and decision theory. With the disquotational acceptance principle meanings are modeled analogously to utilities in decision theory. In decision theory the choice behavior of the subject is determined by her actual utilities together with her probabilistic beliefs. It does not matter what the subject believes about her utilities. But only in exceptional cases where the subject fails to introspect the mechanism behind her own choices she might now know her own utilities.

Let me now explain how the setting of this section, which uses behaviors disquotationally generated from metasemantic models, relates to the simpler setting from Section 2.5, which uses multi-situation models.

First assume that we are given a metasemantic model $M = (W, b, (I_w)_{w \in W})$ and for every situation $s \in S$ an actual world $w_s \in W$. To find a multi-situation model that corresponds to M we need to assume that the semantic facts are the same in all of the situations in which we are interpreting the subject. Formally, this means there is an interpretation function $I : \mathsf{At} \to \mathcal{P}W$ such that $I = I_{w_s}$ for all situations $s \in S$. This condition can be seen as the formal counterpart to the assumption made in Section 2.5 that the language of the subject is the same in all situations. Using this interpretation function we can define the multi-situation model $M' = (W, b, I)$. It is easy to see that this model M' has the property that $a^{M'} = a^M$, where the function on the left side is the behavior generated by the multi-situation model M' according to Definition 2.5.4 and the function on the right side is the behavior disquotationally generated according to Definition 4.2.1.

Conversely, one might also start with a multi-situation model $M = (W, b, I)$ and actual worlds $w_s \in W$ for every situation $s \in S$. Now take any family $(I_w)_{w \in W}$ of interpretation functions with the property that $I_{w_s} = I$ for all $s \in S$. This always exists because we could for instance just set $I_w = I$ for all $w \in W$. For any family of interpretation functions with this property we define the metasemantic model $M' = (W, b, (I_w)_{w \in W})$, which satisfies $a^{M'} = a^M$. This shows that we can choose the interpretation function of any world that is not the actual world of some situation arbitrarily without changing the linguistic behavior that is generated.

The arguments from the previous two paragraphs entail that if we assume that the semantic facts at the actual world of all situations are the same then the behaviors disquotationally generated by some metasemantic model are precisely the behaviors generated by some multi-situation model. And the relation is simply that the interpretation of the multi-situation model is the same as the interpretation at the actual world of the corresponding metasemantic model. Hence metasemantic models and disquotational acceptance do not add anything substantial over the account from Section 2.6. At most we have introduced some new indeterminacy because the semantic facts at any world other than the actual world are not constrained by the linguistic behavior of the subject. This is not a problem since I do not suggest disquotational acceptance and metasemantic models to be an independent account of interpretation. Rather they are one approach for making sense of the interpretation function in a multi-situation model inside the framework of metasemantic models in which semantic facts are explicitly represented. In the next section I introduce another possibility for relating the interpretation function of a multi-situation model with the explicit representation of semantic facts from Section 4.1.

References to the literature

With the disquotational acceptance principle the semantic facts at the actual world determine the linguistic behavior of the subject. Since this principle is so intuitive it seems to me that it is implicitly presupposed by many authors writing on the relationship between meaning and belief.

Kripke (1979, sec. 2) provides an explicit formulation of two principles that together roughly entail the disquotational acceptance principle of this section.

The first, is Kripke's strengthened 'biconditional' disquotational principle. It states that a normal English speaker who is not reticent will be disposed to sincere reflective assent to "p" if and only if he believes that p. Kripke assumes that a similar principle holds for speakers of other languages than English, if we reformulate the whole principle in the language of the speaker. This is necessary because the second occurrence of p in the principle is in the metalanguage which has to match the language of the speaker who accepts the first occurrence of p.

Kripke's second principle is the principle of translation. It states that if a sentence of one language expresses a truth in that language, then any translation of it into any other language also expresses a truth in that other language.

Let us see how one might try to deduce the left-to-right direction of the disquotational acceptance principle from the two principles provided by Kripke. Assume that the subject accepts some sentence "φ". We have then that the sentence "The subject accepts the sentence 'φ'." is a truth in English. We can translate this truth to the language of the subject. Assuming that acceptance amounts is the same as being disposed to sincere reflective assent by a speaker that is not reticent we can then apply Kripke's disquotational principle for the language of the subject to obtain a true sentence in the language of the subject that starts with some words in the language of the subject that mean the same as the English "The subject believes that" and ends with φ outside of quotation marks. This sentence translates to the English sentence that starts with "The subject believes that" and ends with the translation of φ to English. By Kripke's translation principle this English sentence is the also true in English because the original sentence is true in the language of the subject. If we assume that belief is an attitude towards propositions and that the translation of sentences preserves the proposition expressed by the subject then this English sentence entails the claim that the subject believes the proposition that φ expresses in the language of the subject. With similar reasoning, the other way round, one can also derive the right-to-left direction of the disquotational acceptance principle from Kripke's principles.

This suggested derivation of the disquotational principle from Kripke's principles involves some unusual switches between different languages and it makes some assumptions about the semantics of belief ascriptions in English. This difficulties seem to be caused by the fact that Kripke is working in an informal setting whereas the disquotational acceptance principle is formulated in the possible world framework. Nevertheless it seems that the ideas behind Kripke's principles and the disquotational acceptance principle are the same.

4.3 Metasemantic acceptance

In this section I discuss the metasemantic acceptance principle. The metasemantic acceptance principle takes the subject to accept a sentence precisely if she believes that the sentence expresses a true proposition. Thus the connection between the subject's acceptance of sentences and the semantic facts is indirect and goes via the beliefs of the subject.

The disquotational acceptance principle discussed in the previous section identifies the proposition expressed by a sentence that is mentioned in the acceptance principle with the proposition that the sentences expresses according to the semantic facts at the actual world. The metasemantic acceptance principle of this section takes the proposition expressed by a sentence that is mentioned in the acceptance principle to be the proposition that the sentence expresses a true proposition according to the semantic facts. Hence the subject accepts a sentence precisely if in all of her doxastic alternatives the proposition that the sentence expresses according to the semantic facts at the doxastic alternative is true at this doxastic alternative. This leads to the following formulation of the *metasemantic acceptance principle*:

> The subject accepts a sentence if and only if for all of her doxastic alternatives the proposition that the sentence expresses according to the semantic facts that obtain at the doxastic alternative is true at the doxastic alternative.

We can obtain a more concise formulation of the metasemantic acceptance principle if we introduce some new terminology. Let us say that a sentence is true at some possible world if according to the semantic facts that obtain at the world the sentences expresses a proposition that is true at this world. Consequently, the proposition that some sentence is true is the set of all worlds such that the proposition expressed by the sentence according to the semantic facts obtaining at the world is true at the world. In this way we obtain following reformulation of the metasemantic acceptance principle:

> The subject accepts a sentence if and only if she believes that the sentence is true.

As a formal counterpart of the metasemantic acceptance principle we can define the behavior metasemantically generated by a metasemantic model. To see how this works assume that we have a fixed metasemantic model $M = (W, b, (I_w)_{w \in W})$. According to the metasemantic acceptance principle the subject accepts a sentence φ if and only if she believes the proposition that φ expresses a true proposition. Let us formally define the proposition that a sentence φ expresses a true proposition. This proposition is the set of all the worlds w such that the proposition $I_w(\varphi)$ that is expressed by the sentence at the world is true at that world, that is, $w \in I_w(\varphi)$. We can write this proposition as the set $D(\varphi) = \{w \in W \mid w \in I_w(\varphi)\}$. Because of its definition the proposition $D(\varphi)$ is called the *diagonal proposition* expressed by φ in the model M. The subject accepts the sentence φ if and only if she believes the proposition $D(\varphi)$. Thus we obtain the following definition for the behavior metasemantically generated by a model:

4.3.1. Definition. The *behavior* $a^M : S \to \mathcal{PV}$ *metasemantically generated* by a metasemantic model $M = (W, b, (I_w)_{w \in W})$ is defined such that for all situations $s \in S$:

$$a^M(s) = \{\varphi \in \mathcal{V} \mid b(s) \subseteq D(\varphi)\},$$

where $D(\varphi) = \{w \in W \mid w \in I_w(\varphi)\}$.

Consider again the example from Section 4.1 with the domain $W = \{w, v, u, z\}$. It is raining at w and u but not at v and z and at w and v the sentence p means that it is raining whereas at v and z it means that it is not raining. These semantic facts are represented by defining $I_w(p) = I_v(p) = \{w, u\}$ and $I_u(p) = I_z(p) = \{v, z\}$.

Assume that we are interpreting the subject in some situation s where she has the belief set $b(s) = \{w, u\}$. So the subject believes that it is raining and she considers it possible that p means either that it is raining or that it is not raining. In this model we have that $D(p) = \{w, z\}$, which is not a superset of $b(s)$, and hence $p \notin a^M(s)$. Therefore the subject does not accept the sentence according to the metasemantic acceptance principle. Neither does she accept $\neg p$ because $D(\neg p) = \{v, u\}$ is also not a superset of her belief set. The subject neither accepts p nor its negation even though she believes that it is raining. The problem is that she does not know whether she can use p to express this belief.

Next, let me explain how the metasemantic acceptance principle relates to the account of interpretation from Section 2.6. The crucial observation is that the diagonal proposition in a metasemantic model plays the role of the interpretation function in a multi-situation model.

Given a metasemantic model $M = (W, b, (I_w)_{w \in W})$ we can define the multi-situation model $M' = (W, b, D)$, where $D : \mathcal{V} \to \mathcal{P}W$ is the function that maps a sentence φ to its diagonal proposition $D(\varphi)$ in M. To see that this is well-defined in the case where $\mathcal{V} = \mathcal{B}$ on needs to check that the definition of the diagonal proposition satisfies the semantic clauses for the propositional connectives as discussed in Section 1.2. For the model M' one can then show that $a^{M'} = a^M$, where the function on the left side is the behavior generated by M_w according to Definition 2.5.4 and the function on the right side is the behavior metasemantically generated according to Definition 4.3.1.

To define a metasemantic model from a multi-situation model $M = (W, b, I)$ we can use any family of interpretations $(I_w)_{w \in W}$ such that $w \in I_w(\varphi)$ if and only if $w \in I(\varphi)$ for all $w \in W$ and $\varphi \in \mathcal{V}$. Such a family always exists because we could just set $I_w = I$ for all $w \in W$. In general there are however many different such families of interpretations. Given a fixed family of interpretations $(I_w)_{w \in W}$ with this property we can then define the multi-situation model $M' = (W, b, (I_w)_{w \in W})$ for which one can check that $a^{M'} = a^M$.

These considerations show that metasemantic models with the metasemantic acceptance principle yield an account of interpretation that is equivalent to the one from Section 2.6 that uses multi-situation models. I do not think of metasemantic acceptance as giving an independent account of interpretation but rather as a way of making precise how the interpretation function of a multi-situation model relates to the semantic facts.

As with the disquotational acceptance principle there is some indeterminacy if we interpret the subject's behavior with the metasemantic acceptance principle. This is not surprising because metasemantic models are a richer class of structures than multi-situation models. In the case of the disquotational acceptance principle the indeterminacy is that the linguistic behavior of the subject does not constrain the semantic facts at any world of the model that is distinct from the actual world.

The linguistic behavior metasemantically generated by a metasemantic model only depends on the diagonal proposition of the sentences in the language of the subject. So we can not use the linguistic behavior of the subject to distinguish between metasemantic models in which her sentences have the same diagonal proposition. The diagonal proposition of some sentences is defined such that it contains a world if the semantic facts at the world are such that the proposition expressed by the sentence is true at the world. The linguistic behavior of the subject constrains the proposition expressed by some sentence at a doxastic alternative only in so far that it determines whether the proposition is true or false at this doxastic alternative. It does not give us any information to whether the proposition is true at any world that is distinct from the doxastic alternative that we are considering. This shows that to metasemantically generate a behavior from a metasemantic model it suffices that sentences express truth values at the worlds of a model. There is no need to let sentences express propositions.

We can view the choice between disquotational and metasemantic acceptance as a choice between two different trade-offs for setting up our formal models. With the disquotational acceptance principles sentences need to express propositions but we only need to know the proposition expressed at the actual world. With the metasemantic acceptance principle sentence just need to have truth values but we need to know the truth value of a sentence at every world that is a doxastic alternative for the subject in some situations.

In Chapter 6 I extend the account of interpretation such that we can assume that the language of the subject contains modalities that operate on the proposition expressed by a sentence. In this setting the full proposition that is expressed by a sentence at some doxastic alternative is relevant to determine the behavior metasemantically generated by some metasemantic model.

References to the literature

The metasemantic acceptance principle is based on an idea developed by Stalnaker (1978). The problem that Stalnaker addresses is different from the one of this thesis. Let me explain how the two relate.

Stalnaker (1978) is concerned with determining what information an assertion adds to the common ground of a conversation, if it is accepted by all participants of the conversation. The common ground of a conversation is all the information that is presupposed by the participants of the conversation. Stalnaker models the common ground by a set of possible words that contains all the worlds which are compatible with what is presupposed by the participants of the conversation. The common ground of a conversation is modeled similarly to the belief set of some subject, which is a set of all the world that are compatible with everything that the subject believes.

To determine what information an assertion adds to the common ground of a conversation Stalnaker first determines what information the assertion conveys relative to the common ground. He assumes that this information is represented by a proposition, which is just a set of worlds that can then be added to the common ground by taking the intersection. Stalnaker also assumes that the sentence uttered in the assertion expresses a proposition relative to every world in the common ground. To model this he uses the framework that I present in Section 4.1. Stalnaker's problem of determining the information conveyed by an assertion is thus analogous to the problem of this chapter which is to make precise how to determine the proposition expressed by a sentence that is mentioned in the acceptance principle from Section 1.3.

Stalnaker first considers the case where the asserted sentences expresses the same proposition relative to all worlds in the common ground. In this case the information conveyed by the sentence is simply this proposition that the sentence expresses relative to any of the worlds in the common ground.

It is more interesting to account for cases where the proposition expressed by a sentence is different in at least two worlds in the common ground. Stalnaker suggests that in such cases people commonly take the diagonal proposition to be the information conveyed by a sentence. This corresponds to the metasemantic acceptance principle discussed in this section.

Stalnaker describes the use of the diagonal proposition to obtain the information conveyed by an assertion as a reinterpretation that is only applied in some cases where the common ground does contain worlds that do not agree on the semantic facts relevant for the interpretation of the assertion. If one uses the metasemantic acceptance principle then one takes this reinterpretation to be the standard interpretation that is used in all cases.

It might seem as if Stalnaker suggestion were much more modest than the metasemantic acceptance principle. However, note that if a sentence expresses the same proposition in all the worlds of the common ground then at every world in the common ground this proposition has the same truth value as the diagonal proposition. Therefore the result obtained by intersecting the common ground with the proposition expressed by the sentence in some world of the common ground is the same as result obtained by intersecting it with the diagonal proposition. Hence, if Stalnaker made the slightly stronger claim that reinterpretation with the diagonal proposition applies in all cases where the worlds on the common ground do not agree on the proposition expressed by the sentence then his theory would be equivalent to one that interprets all assertions with the diagonal proposition. Thus, he would be using a counterpart of the metasemantic acceptance principle.

Stalnaker (1978, p. 327) also discusses the possibility of taking the information conveyed by the assertion of a sentence to be the proposition the sentence expresses according to the actual world. This is the counterpart of disquotational acceptance principle in his setting. He notices that he can not use this solution because it often is unknown to the participants of the conversation what the actual world is. Hence they would not be able to add the proposition expressed according to the actual world to the common ground. I think that on this point Stalnaker is somewhat mislead by the power of his own framework, which is designed to represent uncertainty about the semantic facts. If one uses disquotational acceptance then, as I explain in Section 4.2, the cases where the subject does not know the relevant semantic facts are exceptional cases in which she fails to introspect on her own linguistic behavior.

In the literature there is generally quite some suspicion to the view that the subject's own beliefs about the semantic facts factor into the proposition that she expresses with her utterances. As an example Burge (1979, secs. IIIb and IIIc) extensively argues against what he calls the metalinguistic reinterpretation of the subject's utterances on the basis that it does not conform to common practice.

4.4 The status of the distinction

In this section I am concerned with the theoretical status of the distinction between disquotational and metasemantic acceptance. I suggest that the distinction is mainly theoretical in that it regulates the formal structure of the account of interpretation but has no or very little effect on the class of interpretable behaviors.

The disquotational and metasemantic acceptance principles provide different ways of understanding the proposition expressed by a sentences that is used in the acceptance principle from the first part of this thesis. In Sections 4.2 and 4.3 I explain how they connect the interpretation of a multi-situation model differently to the semantic facts that are represented in a metasemantic model. But since both of them relate back to multi-situation models they do not lead to an account of interpretation that is different from the one given in Section 2.6.

One might wonder why we should bother with the difference between the disquotational and the metasemantic acceptance principles. They just provide distinct reformulations of the account from Section 2.6 in the more complex setting of metasemantic models.

In the following two chapters of this thesis I am considering problems that motivate representations of meaning that are more complex than the interpretation function of a multi-situation model. The distinction between disquotational and metasemantic acceptance influences the mathematical structure of these more complex representations of meaning and it suggests different ways of thinking about these structure. In Chapter 5 I show that the metasemantic acceptance principle is presupposed by an explanation for the splitting of possible worlds from Section 2.4. With disquotational acceptance it is more natural to avoid the splitting of worlds and use a more complex representation of meaning instead. In Chapter 6 I show that depending on which acceptance principle one chooses we obtain different formal semantics for a necessity modality in the language of the subject.

Let us now assume that the discussion in the following chapters indeed shows that depending on which version of the acceptance principle one uses it is natural to employ different formal models to account for the same linguistic phenomena. In this case it becomes interesting to find a reasons for preferring one of the acceptance principles over the other because this would then also give us a reason for preferring one type of formal model over another. What kind of reason could that be?

Ideally, it would turn out that the natural account of interpretation for one of the acceptance principles performs better on the requirements from Section 1.4 than the natural account of interpretation for the other acceptance principle. This would give us a reason to prefer the former acceptance principle.

In this thesis I do not reach the point where one of the acceptance principles leads to an account of interpretation that performs decisively better than the account for the other principle. Hence I leave it open which principle to choose and I content myself with investigating the differences between them.

The argumentative structures in Chapters 5 and 6 are similar. In both chapters it initially seems that metasemantic acceptance performs better than disquotational acceptance. But in both cases disquotational acceptance can be improved by introducing a more complex notion of semantic facts. There is a reason for this common pattern that I want to bring to the reader's attention because it explains why it is difficult to turn the distinction between disquotational and metasemantic acceptance into a difference between the classes of interpretable behaviors.

With disquotational acceptance the semantic facts at the actual world determine the linguistic behavior of the subject. With metasemantic acceptance the beliefs of the subject about the semantic facts determine her linguistic behavior. The difference between the approaches is that metasemantic acceptance incorporates the complexity of the doxastic structure into how the structure representing meanings determines linguistic behavior whereas disquotational meaning does not. This detour over the beliefs of the subject is an advantage because in many cases it turns out that this additional complexity of the doxastic structure helps to account for plausible linguistic behaviors. Hence it seems initially as if metasemantic acceptance can account for a richer class of behaviors than disquotational acceptance. But we can fix disquotational acceptance by including the relevant feature of the doxastic structure into the notion of semantic facts. With this enhanced notion of semantic facts the richer semantic structure associated to a sentence at the actual world behaves similar to the original simpler semantic structure distributed over the doxastic alternatives of the subject. As a consequence the class of interpretable behaviors become the same.

Let me give a simple example in which this pattern already occurs. For this we consider again the question whether the semantic facts should be relativized to the language of the subject. Assume that we want to interpret two subjects X and Y that speak two different languages that share a common vocabulary. This might happen with two speakers of a single community that attach their own idiosyncratic meanings to certain words, or it could be that they belong to different linguistic communities that by coincidence use the same sentences. Suppose that X and Y do not interact, and hence we need not model the beliefs that they have about each other. But we do want to interpret both of them with a single metasemantic model. For the doxastic part of the model we need two belief functions, b_X for X and b_Y for Y. This is common practice in multi-agent doxastic logics because obviously different people can have different beliefs. But do we also need two families of interpretation functions, one for X and one for Y? The answer depends on our choice of the acceptance principle.

With metasemantic acceptance there is no need to relativize the family of interpretation functions to the subjects. We can use metasemantic models of the form $(W, b_X, b_Y, (I_w)_{w \in W})$. That X and Y are speaking different languages is represented in such a model by them having different beliefs about the semantic facts. The interpretation functions that are associated to the doxastic alternatives of X are different from the interpretation functions that are associated to the doxastic alternatives of Y. There is no need to relativize the interpretation functions in the model because the relativization of the doxastic structure already accounts for the difference in meanings that influences linguistic behavior.

With disquotational acceptance the interpretation function of the actual world is used to interpret the behavior of some subject. Hence we need to relativize the interpretation functions to subjects because we want to use a different interpretation for the sentences used by X than we use for the sentences used by Y. This yields metasemantic models $(W, b_X, b_Y, (I_{X,w})_{w \in W}, (I_{Y,w})_{w \in W})$ that contain two families of interpretations, $(I_{X,w})_{w \in W}$ to represent meanings in the language of X and $(I_{Y,w})_{w \in W}$ to represent meanings in the language of Y.

Chapter 5

Semantic indeterminacy or uncertainty

This chapter concerns the problem that in some situations the number of worlds in the belief set of the subject does not suffice to account for her acceptance of propositional sentences. I am calling this the problem of vagueness because cases involving vague expressions are typical instances of the problem, in which the subject does not accept some sentence nor its negation even though she has complete information about all the relevant basic facts.

Solving the problem of vagueness brings us back to the issues of splitting possible worlds, which is introduced in Section 2.4 as a technical trick without much justification. In Section 5.1 I show that if one accepts the metasemantic acceptance principle then the splitting of possible worlds can be understood in the setting from Chapter 4 as introducing uncertainty for the subject about the semantic facts. This explanation does not work well with disquotational acceptance and hence it is left open how to account for vagueness in this case.

In Section 5.2 I solve the problem of vagueness in the case of disquotational acceptance by representing meanings with a set of interpretation functions instead of a single interpretation functions. The resulting account of interpretation renders exactly those linguistic behaviors interpretable that are also interpretable using the trick of splitting possible worlds.

In the last section of this chapter I explain that there is still an important conceptual difference between the account that uses an interpretation over a splitting of the domain and the account that uses a set of interpretation functions. On the former but not the latter semantic information can be entangled with information about the basic facts. It however turns out that for the simple accounts of interpretation discussed here this does not influence the class of interpretable behaviors.

I want to stress that this chapter does not aim at providing a state-of-the-art formal semantics for vague expressions. It rather concerns a problem internal to the theory of interpretation that arises whenever the uncertainty in the belief set of the subject is not enough to account for her reluctance to accept propositional sentences. In general we might think of such cases as cases in which either the subject has additional uncertainty about the semantic facts or in which the semantic facts themselves are indeterminate. The formal techniques that are used in this chapter to solve this general problem have been used in the literature specifically to give a semantics of vague expressions.

5.1 Splittings for metasemantic acceptance

I start this section by recalling the solution to the problem of vagueness from Section 2.4, which is solved by splitting possible worlds. I then explain how we can make sense of this splittings if we are using the metasemantic acceptance principle.

The problem of vagueness arises for the notion of tight interpretability that is introduced in Section 2.3 and then adapted in Section 2.6 to the case in which we interpret the subject across multiple situations. I sketch the notion of tight interpretability in the following paragraphs, but for the details the reader is referred to these sections and especially to Definition 2.6.1.

Tight interpretability is probably the most intuitive notion of interpretability for an account of interpretation that presupposes prior knowledge about the beliefs of the subject. For tight interpretability the interpretation function that interprets the behavior of the subject needs to be defined on the domain of possible worlds that has been fixed in advance with the purpose of encoding our prior knowledge about the beliefs of the subject.

The difficulty with tight interpretability arises because by requiring that the interpretation that represents meanings in the language of the subject is defined over the domain that is fixed in advance we might not have enough worlds to make sense of her linguistic behavior. It is shown in Theorem 7.3.10 that tight interpretability entails the cardinality condition from Definition 2.6.9. This condition requires that the set of sentences $a(s)$ that the subject accepts in some situation s satisfies a special constraint that depends on the number of elements in her belief set $b(s)$ in this situation. In the special case where $b(s)$ is a singleton set, so the subject has no uncertainty about the relevant basic facts, these constraints amount to the requirement that for every sentence the subject accepts it or its negation.

One can use sentences that involve vague expressions to give examples of linguistic behaviors that show that the conditions for tight interpretability are too strong. Let me recall the example from the end of Section 2.4. There we assume that the subject is a speaker of English and so her language contains the sentence "The man is tall." for which we use the letter p, and it contains the sentence "The man is not tall." which is the negation $\neg p$ of p. We consider the domain $W = \{w_{1.6}, w_{1.8}, w_{2.0}\}$ such that at w_h the man that is the referent of "The man" in the sentence p is h meters tall. The subject is in a situation s, where her belief set is the singleton $b(s) = \{w_{1.8}\}$, so she is fully informed about the relevant basic facts. Nevertheless she neither accepts p nor $\neg p$, which we can represent by $a(s) = \mathsf{cl}\,(\emptyset)$, saying that in s she only accepts the tautologies in the propositional language generated by p. This behavior does not satisfy the constraints for tight interpretability, and hence tight interpretability does not fulfill the variety requirement.

It might be that there are also examples that do not involve vague expressions that are counterexamples against the special constraints on tight interpretability. If this is the case then the problem that I deal with in this chapter is broader than just the problem of vagueness. I however focus on the problem vagueness because the technical approaches that I use for solving the problem are commonly treated in the context of vagueness.

In Section 2.4 I develop an account of interpretations that does not impose the strong constraints on the behavior of the subject which cause the problem of vagueness for tight interpretability. This leads to the notion of splitting interpretability which is defined in Definition 2.6.2.

Splitting interpretability allows that the interpretation function that interprets the behavior of the subject is defined over a split domain that is larger than the original domain on which the belief function of the subject is defined. This split domain needs to be connected to the original domain with a surjective function which I call the splitting function. When a world in the split domain is mapped by the splitting function to a world in the original domain this tells us that the same basic facts obtain at the world in the split domain and at the world in the original domain. As a consequence every proposition relative to the original domain corresponds to a proposition relative to the split domain that encodes the same information about the basic facts and is given by the preimage under the splitting function. This is useful since it allows us to transfer the belief sets of the subject that are defined on the original domain to the split domain on which the interpretation function of the interpreting model is defined.

Let me explain on our example how splitting interpretability deals with the problem of vagueness. The trick is to split the world $w_{1.8}$ such that at one of its copies the sentence p is true but at the other copy the sentence is false. For this we use the split domain $W' = \{w_{1.6}, w_{1.8}, w'_{1.8}, w_{2.0}\}$ with a splitting function $f : W' \to W$ that maps w_h on w_h and $w'_{1.8}$ on $w_{1.8}$.

In the situation where we are interpreting the subject she believes that the man that is referred to in p is 1.8 meters tall. So her belief set relative to W is $\{w_{1.8}\}$, which corresponds to the belief set $\{w_{1.8}, w'_{1.8}\}$ relative to W'. An interpretation function $I : \mathsf{At} \to \mathcal{P}W'$ over W' that interprets the behavior of the subject in this example must be such that p is true at one of $w_{1.8}$ and $w'_{1.8}$ and false at the other. For instance it might be such that $I(p) = \{w_{1.8}, w_{2.0}\}$. We can collect this data into one splitting interpretation model $M = (W', f, I)$ for which one can easily see that it interprets the linguistic behavior a with the belief function b.

In Section 2.4 I introduce the splitting of worlds as a technical trick that allows us to get rid of the strong constraints on tight interpretability. In this section I explain that with the metasemantic acceptance principle the splitting of worlds can be understood as introducing the uncertainty that the subject has about the semantic facts.

Let me first show on our example how this works and then explain the general idea.

The difference between the worlds $w_{1.8}$ and $w'_{1.8}$ lies in the semantic facts that obtain at these worlds. The semantic facts at $w_{1.8}$ are such that the sentence p expresses a proposition that is true at $w_{1.8}$. At $w'_{1.8}$ on the other hand the semantic facts are such that the sentence p expresses a proposition that is false at $w'_{1.8}$. In this way the truth value of the sentence p at $w_{1.8}$ can be different from its truth value at $w'_{1.8}$ even thought the same basic facts obtain at both worlds.

One can make this difference in semantic facts more explicit by providing a metasemantic model of the kind discussed in Section 4.1. One model that would fit nicely in the example uses the family of interpretation functions that is given in the following table:

	$w_{1.6}$	$w_{1.8}$	$w'_{1.8}$	$w_{2.0}$
$w_{1.6}$	$\neg p$	p	p	p
$w_{1.8}$	$\neg p$	p	p	p
$w'_{1.8}$	$\neg p$	$\neg p$	$\neg p$	p
$w_{2.0}$	$\neg p$	p	p	p

The semantic facts represented in this table are such that according to the semantic facts at $w_{1.6}$, $w_{1.8}$ and $w_{2.0}$ a man of height 1.8 already counts as tall whereas according to the semantic facts holding at $w'_{1.8}$ he does not. Hence at $w_{1.6}$, $w_{1.8}$ and $w_{2.0}$ the sentence p expresses a proposition that is true at all the worlds where the basic facts are such that the man is at least 1.8 meters tall whereas at $w'_{1.8}$ the proposition expressed by p is true only at $w_{2.0}$.

One could also consider copies $w'_{1.6}$ of $w_{1.6}$ and $w'_{2.0}$ of $w_{2.0}$ at which the same semantic facts obtain as at $w'_{1.8}$. In the discussion of the example these worlds are never needed as doxastic alternatives of the subject and hence I have simplified the model by not including them.

Because metasemantic models explicitly represent the semantic facts that hold at different worlds it is convenient to use them here to illustrate an uncertainty about the semantic facts. However, as explained at the end of Section 4.3, it however suffices to work with multi-situation model when defining the notion of splitting interpretability. I could also represent the example in a multi-situation model, whose interpretation function is the diagonal of the metasemantic model given here.

In a situation where the subject has the belief set $\{w_{1.8}, w'_{1.8}\}$ she is uncertain about the semantic facts. She does not know whether "tall" applies to men of height 1.8 or not. Consequently she does not know what proposition is expressed by the sentence p. She does also not believe that the proposition expressed by p is true. According to her doxastic alternative $w_{1.8}$ it is and according to the doxastic alternative $w'_{1.8}$ it is not. Similarly she does not belief that the proposition expressed by $\neg p$ is true. According to $w'_{1.8}$ it is but according to $w_{1.8}$ it is not.

With the metasemantic account of acceptance the uncertainty that the subject has about the semantic facts influences her linguistic behavior. The subject accepts a sentence if and only if she believes that it expresses a proposition that is true. In the example the subject neither believes that p expresses a true proposition nor does she believe that $\neg p$ expresses a true proposition. Hence she accepts neither p nor $\neg p$.

Let me now explain in more general terms how to understand the splitting of worlds that happens with splitting interpretability.

The general structure of a splitting $f : W' \to W$ should be understood as follows: The original domain W contains all relevant combination of basic facts, whereas the split domain W' contains all relevant combination of basic facts and semantic facts. The splitting function reduces a combination of basic and semantic facts to its part that is a combination of basic facts and forgets about the semantic facts.

When we specify the beliefs of the subject in advance to interpretation relative to the domain W then we only specify her beliefs about the basic facts. If her belief set is a singleton subset of W then she has complete information about all the basic facts. It does however not follow that she also has complete information about the semantic facts. In cases of vagueness her belief set, considered as a subset of W', is not a singleton set.

On the metasemantic account of acceptance it is part of the process of interpretation to discover the beliefs that the subject has about the semantic facts. This is why we start from the domain W that does not represent variations in semantic facts and then interpret the subject with a domain W' that includes variations in semantic facts.

With metasemantic acceptance the subject accepts a sentence precisely if she believes that according to the semantic facts it expresses a proposition that is true. Her beliefs about the basic facts together with her beliefs about the semantic facts determine whether she accepts a sentence. This is why in cases of vagueness the subject accepts neither a sentence nor its negation even though she has complete information about the basic facts. In such cases she is uncertain about the semantic facts.

The explanation of the splitting as introducing the subject's uncertainty about the semantic facts is not convincing when we are using the disquotational acceptance principle. Let me explain this in detail.

When we split possible worlds to interpret the subject then we consider her uncertainty about some facts that are distinct from the basic facts that are represented by the original domain. This general explanation is also available with the disquotational acceptance principle. But if we use disquotational acceptance it is difficult to say what kind of facts these additional uncertainty is about.

With the metasemantic acceptance principle it is plausible to take the additional uncertainty to be about the semantic facts because these have a direct influence on the linguistic behavior of the subject. With the disquotational acceptance principle this is not the case. The linguistic behavior of the subject depends on the semantic facts obtaining at the actual world. When we interpret the subject we discover the semantic facts that obtain at the actual world. In a case of vagueness we discover that the proposition expressed by some sentence at the actual world is sensitive to the additional facts that are represented in the split domain. If we take these additional facts in the split domain to be semantic facts then this entails that the meaning of sentences at the actual world are sensitive to these semantic facts. But it is implausible that in cases of vagueness the sentences of the subject are about semantic facts.

Let us consider again the example from above to illustrate this point. In the model that interprets the linguistic behavior of the subject there are two possible worlds $w_{1.8}$ and $w'_{1.8}$ which correspond to the same combination of basic facts but are distinct with respect to some of the additional facts that the splitting has introduced. The interpretation function of the interpreting model is such that the proposition expressed by p is true at $w_{1.8}$ and false at $w'_{1.8}$. With the disquotational acceptance principle we think of this interpretation function as encoding the semantic facts that obtain at the actual world. The proposition that is expressed by p according to the semantic facts at the actual world is sensitive to the facts that distinguish between $w_{1.8}$ and $w'_{1.8}$. If we take these distinguishing facts to be semantic facts then the sentence p itself is at least partially about the semantic facts. But we took the example to be such that the subject is a speaker of English and p corresponds to the sentence "The man is tall." It is rather implausible that when English speakers use "The man is tall.' they are saying something about the semantic facts.

I do not take the considerations from the previous paragraph to show that with disquotational acceptance there is absolutely no way of making sense of the splitting of possible worlds. However in the following section I discuss an account of interpretation that solves the problem of vagueness without splitting possible worlds and hence combines better with disquotational acceptance.

References to the literature

Accounting for the splitting of worlds as introducing uncertainty about the semantic facts is a variant of epistemicism that has been defended by Williamson (1994, chs. 7 and 8) as a solution to the problem of vagueness. It is not clear to me how the account from this section relates precisely to the account of Williamson. Williamson claims that in cases of vagueness the subject is uncertain about the relevant fact. For instance in the case discussed above that involves the vague sentence “The man is tall.” the subject would be uncertain about whether the man is tall, even though she knows his precise height. In the account of this section there is no need to postulate a basic fact such as the tallness of the man. Instead the subject is uncertain about the semantic facts that determine the meaning of the sentence “The man is tall.”

In his discussion Williamson however seems to assume that in cases of vagueness the subject’s uncertainty about facts such as that the man is tall is a consequence of her uncertainty about the use of the relevant concepts. He also suggests that there is a close connection between the use of a term and its meaning. So it seems that also on Williamson’s view the subject’s uncertainty in cases of vagueness is at least partially about the semantic facts.

5.2 Supervaluations for disquotational acceptance

In this section I consider an account of interpretation that avoids the undesirable constraints for tight interpretability without needing to split possible worlds. This provides a solution to the problem of vagueness that is more attractive than splitting interpretability if one uses the disquotational acceptance principle.

The idea behind the account of this section is to take cases of vagueness to be cases in which the semantic facts do not determine whether a sentence is true or false at the world that is the singleton belief set of the subject in a situation where she has complete information about the basic facts. This requires us to represent the semantic facts with some structure that is more complex than an interpretation function because according to an interpretation function every sentence is either true or false at every world.

The most straightforward adaption of the notion of an interpretation function on which it can be indeterminate whether a sentence is true or false at some world is to use interpretation functions that map sentences to multi-valued propositions that are functions from worlds to some set of truth-values with more than two elements. Such an approach, although straightforward, does not fit well into the setting of this thesis because it gives rise to logics that are not classical.

In this section I use a set of interpretation functions to represent the semantic facts that obtain at some world. The truth value of some sentence can then be indeterminate in the sense that it is true relative to some and false relative to some other interpretation function in the set. Such sets of interpretation functions are called supervaluations:

5.2.1. Definition. A *supervaluation $\mathcal{I}$ over W* is a non-empty set of interpretation functions such that every $I \in \mathcal{I}$ is a function $I : \mathsf{At} \to \mathcal{P}W$ that maps atomic sentences to propositions over W.

Let us consider in an example how a supervaluation represents the indeterminacy of semantic facts. We are using the domain $W = \{w_{1.6}, w_{1.8}, w_{2.0}\}$ in which the basic facts are such that at w_s the man is s meters tall. Let p be the sentence "The man is tall." The semantic fact that in English the sentence this sentence is neither true or false of a man that is 1.8 meters tall is represented for instance by the supervaluation $\mathcal{I} = \{I, I'\}$ over W such that $I(p) = \{w_{1.8}, w_{2.0}\}$ and $I'(p) = \{w_{2.0}\}$.

To obtain an account of interpretation that uses supervaluations we need to alter the definitions of possible world models to include these supervaluations instead of just an interpretation function. Thus we obtain the following analogue of Definition 2.5.3 for multi-situation possible world models:

5.2.2. Definition. A *supervaluation model* $M = (W, b, \mathcal{I})$ is a domain W together with a belief function $b : S \to \mathcal{P}W$ and a supervaluation $\mathcal{I}$ over W.

An example of a supervaluation model is the model $M = (W, b, \mathcal{I})$, where $W = \{w_{1.6}, w_{1.8}, w_{2.0}\}$ is the domain from the paragraph above, $b : S \to \mathcal{P}W$ is a belief function that maps a single situation s to the belief set $b(s) = \{w_{1.8}\}$, and $\mathcal{I} = \{I, I'\}$ is the supervaluation from the paragraph above. This model captures a situation where the subject has complete information about the basic facts but her language leaves it indeterminate whether p applies in that situation.

One might also adapt Definition 4.1.1 to obtain a supervaluation analogue of metasemantic possible world models. To do so one replaces the family $(I_w)_{w \in W}$ of interpretation functions with a family $(\mathcal{I}_w)_{w \in W}$ of supervaluations. This family $(\mathcal{I}_w)_{w \in W}$ specifies for every possible world w the supervaluation $\mathcal{I}_w$ that represents the semantic facts obtaining at w. In this thesis I do not need the definition of metasemantic supervaluation models because I use supervaluations only in combination with a disquotational acceptance for which it suffices to have only one supervaluation that represents the semantic facts that obtain at the actual world.

With supervaluations there is no such thing as the proposition that a sentence expresses according to the semantic facts at some world. A sentence might expresses a different proposition relative to the different interpretation functions that are in the supervaluation associated to the world. There is no unique proposition that is expressed by a sentences. Because the notion of the proposition expressed by some sentence is used in the acceptance principles defined so far we can not apply these acceptance principles to the setting of supervaluations. We need to define a new acceptance principle that suits supervaluation models.

The most natural adaption of the disquotational acceptance principle is to universally quantify over all the interpretation functions in the supervaluation of the actual world. Thus we obtain the following *disquotational acceptance principle for supervaluations*:

> The subject accepts a sentence if and only if she believes all of the

propositions that the sentence expresses according to the semantic facts that obtain at the actual world.

The formal counterpart of the disquotational acceptance principle for supervaluation is the following notion of the behavior generated by a supervaluation model:

5.2.3. Definition. The *linguistic behavior* $a^M : S \to \mathcal{PV}$ *generated by a supervaluation model* $M = (W, b, \mathcal{I})$ is defined such that for all $s \in S$

$$a^M(s) = \{\varphi \in \mathcal{V} \mid b(s) \subseteq I(\varphi) \text{ for all interpretations } I \in \mathcal{I}\}.$$

For this definition to be a formalization of the disquotational acceptance principle for supervaluations one needs to think of the supervaluation $\mathcal{I}$ in the supervaluation model $M = (W, b, \mathcal{I})$ as representing the semantic facts of the actual world.

One could also define a metasemantic acceptance principle for supervaluations by requiring for acceptance of some sentence that in every doxastic alternative of the subject all of the proposition that the sentence expresses according to the semantic facts that obtain at that doxastic alternative are true at the doxastic alternative. I am not pursuing this approach here because I am using supervaluations as a solution to the problem of vagueness for an account that is based on disquotational acceptance. Combining supervaluations with the metasemantic acceptance principle would be unnecessary complicated since splitting interpretability already solves the problem of vagueness for metasemantic acceptance.

I continue by defining a notion of interpretability for supervaluations.

5.2.4. Definition. A supervaluation $\mathcal{I}$ over W *interprets a linguistic behavior* $a : S \to \mathcal{PV}$ *with a belief function* $b : S \to \mathcal{P}W$ if $a = a^M$ for the supervaluation model $M = (W, b, \mathcal{I})$.

A linguistic behavior $a : S \to \mathcal{PV}$ is *supervaluation interpretable with a belief function* $b : S \to \mathcal{P}W$ if there exists some supervaluation over W that interprets a with b.

Definition 5.2.4 is an adaption of tight interpretability for multi-situation models, as defined in Definition 2.6.1, to supervaluations. Especially it does not involve any splitting of possible worlds as is needed for the notion of splitting interpretability from Definition 2.6.2.

Let us consider again the supervaluation model $M = (W, b, \mathcal{I})$ from above, where $W = \{w_{1.6}, w_{1.8}, w_{2.0}\}$, $b(s) = \{w_{1.8}\}$ and $\mathcal{I} = \{I, I'\}$ contains the interpretations $I, I' : \mathsf{At} \to \mathcal{P}W$ with $I(p) = \{w_{1.8}, w_{2.0}\}$ and $I'(p) = \{w_{2.0}\}$. This model plausibly represents the case of vagueness discussed in this chapter. One can see that if p is the only atomic sentence in the language of the subject then the behavior generated by the supervaluation model M is such that $a^M(s) = \mathsf{cl}(\emptyset)$. According to the model M the subject accepts neither p nor $\neg p$ in the situation

s. Hence the linguistic behavior in the example of vagueness is supervaluation interpretable.

I continue by evaluating supervaluation interpretability as an account of interpretation and comparing it to the notion of splitting interpretability from Section 2.6.

I start with the variety requirement which is that all linguistic behaviors that a subject might plausibly show should be interpretable. For this we need to know which behaviors are supervaluation interpretable. Theorem 7.5.1 shows that a linguistic behavior is supervaluation interpretable with some given belief function if and only if it splitting interpretable with this belief function. Hence the class of supervaluation interpretable behaviors is precisely the same as the class of splitting interpretable behaviors. This shows that as far as the variety requirement is concerned the two accounts of interpretation are equivalent.

For a precise characterization of the supervaluation interpretable behaviors we can use Theorem 7.3.4 which shows that some behavior is splitting, and hence by Theorem 7.5.1 supervaluation, interpretable with some belief function precisely if it satisfies the conjunctive covering condition and the conjunctive consistency condition. I already explain these conditions in Section 2.6 for the notion of splitting interpretability. For the discussion here it is just relevant that supervaluation interpretability does not impose the strong cardinality constraints of tight interpretability. Thus it avoids the problem of vagueness.

On the little-input requirement supervaluation interpretability also performs equally well as splitting interpretability. Both accounts presuppose that for every situation we know the subject's belief set and that we have a hypothesis about which constructions in the language of the subject play the role of the logical connectives. I consider especially the former assumption as being too strong. One might weaken this to only assuming that we know for every situation what evidence the subject has and then develop an account of interpretation with that uses supervaluations and plausibility orders analogous to the account from Section 3.3. I do not develop such an account in this thesis but I have some remarks about it in Section 5.3 where I show that in the setting of plausibility orders supervaluation and splitting interpretability might differ with respect to the variety requirement.

Lastly, there might be differences between splitting and supervaluation interpretability with respect to the determinacy requirement because meanings are represented by different structures. I have not looked carefully enough at the problem of determinacy for either account to find such differences. Neither notion of interpretability suffers from an obvious indeterminacy that renders it unusable.

References to the literature

Supervaluations are widely used to model semantic indeterminacy. Van Fraassen (1966) uses them to account for sentences that contain singular terms that fail to refer while retaining classical logic. An application of supervaluations to the problem of vagueness is given by Kamp (1975) and Fine (1975). The formal presentation of the framework in these papers is more complex than the one given in this thesis. They are however equivalent for the application of this thesis in which the acceptance of a sentence is determined by universal quantification over all the classical interpretation functions in a supervaluation.

5.3 Entanglement

In this section I discuss an important difference between supervaluation interpretability, based on disquotational acceptance, and splitting interpretability, based on metasemantic acceptance. In metasemantic models the subject's beliefs about the semantic fact can be entangled with her beliefs about basic facts. This is not the case in supervaluation models in which the indeterminacy of meanings that is represented by the supervaluation is independent from the uncertainty about the basic facts that is represented by the belief set. I give an example in which it is quite intuitive to represent the subject as having beliefs about meanings that are entangled with her beliefs about the basic facts. Nevertheless I explain at the end of this section that it is not clear how the possibility to represent entanglements influences the formal account of interpretation.

Under entanglement I understand the property that the beliefs of the subject about the semantic facts can depend on her beliefs about the basic facts and, vice versa, that her beliefs about the basic facts can depend on her beliefs about the semantic facts. Let us consider an example. We use a domain $W = \{w_0, w_1, v_0, v_1\}$ with four possible worlds to represent the basic facts. One basic fact is whether it is raining. It is raining at w_0 and w_1 but not at v_0 and v_1. The second basic fact is whether some other person that the subject is observing is using the sentence p or the sentence $\neg p$ to describe the weather. In w_1 and v_1 this other person is accepting the sentence p and in w_0 and v_0 he is accepting the sentence $\neg p$.

Assume that we are in a situation s where the subject does not know whether it is raining and she does not know whether the other person is accepting p or $\neg p$. Hence her belief set in s contains all of the worlds in the original domain W, or at least splitting copies of those. Let us also suppose that the subject is a child that is adapting her beliefs about meanings to the those of the people around her. She does not know whether p means that it is raining or whether it means that it is not raining. But she believes that the other person is uttering a true sentences. It follows for instance that at the doxastic alternative w_0, where it is raining and the other person is accepting $\neg p$, the sentence p means that it is not raining. Or, at w_1, where it is also raining but the other person is accepting p, the sentence p means that it is raining. Similarly, one can determine the meanings of p at the other worlds v_0 and v_1.

One can represent this beliefs of the subject about the semantic facts with metasemantic model $(W, b, (I_w)_{w \in W}))$ in which $b(s) = W$ and the semantic facts are such that $I_{w_1}(p) = I_{v_0}(p) = \{w_0, w_1\}$ and $I_{w_0}(p) = I_{v_1}(p) = \{v_0, v_1\}$. At w_1 and v_0 the sentence p means that it is raining, whereas at w_0 and v_1 it means that it is not raining. This family of interpretations is presented more conveniently in the following table:

	w_0	w_1	v_0	v_1
w_0	$\neg p$	$\neg p$	p	p
w_1	p	p	$\neg p$	$\neg p$
v_0	p	p	$\neg p$	$\neg p$
v_1	$\neg p$	$\neg p$	p	p

Now imagine that the subject learns that the other person accepts the sentence p. So we are in a new situation t, where she has the belief set $b(t) = \{w_1, v_1\}$. In this situation the subject still does not know whether it is raining and she still does not know whether p means that it is raining or whether it means that it is not raining. But she knows that p means that it is raining exactly if it is raining and that p means that it is not raining exactly if it is not raining. Hence her beliefs about the basic facts and her beliefs about meanings are entangled. If the subject were to learn that it is raining, so her belief set would shrink to the singleton $\{w_1\}$, she would also start to believe that p means that it is raining. And, vice versa, if the subject were to learn from somewhere else that p means that it is raining then she would thereby also come to believe that it is raining.

The kind of entanglement of beliefs about basic facts with beliefs about semantic facts that is discussed in the preceding example can be represented with metasemantic possible world models. Because we can combine metasemantic models with both disquotational and metasemantic acceptance it might seem that the issue of representing entanglement is independent of whether we use the disquotational or the metasemantic acceptance principle. Nevertheless I claim that entanglement is characteristic of metasemantic acceptance. On the metasemantic acceptance principle the beliefs that the subject has about the semantic facts influence her linguistic behavior. Hence it can have an effect on her linguistic behavior if her beliefs about semantic facts are entangled with her beliefs about the basic facts. On the disquotational acceptance principle we are using the semantic facts of the actual world to interpret the subject. The semantic facts in her belief set, which might be entangled with the basic facts, are irrelevant for her linguistic behavior.

Consider again the situation t in the example where the subject's beliefs about the meaning of p are entangled with her beliefs about the weather. Suppose that now the subject learns that it is raining and so we are in a new situation z, where her belief set is the singleton $\{w_1\}$. According to the metasemantic acceptance principle the subject would start accepting the sentence p in this situation z because she believes that p means that it is raining and she believes that it is raining. On the disquotational acceptance principle there would be no such connection between the semantic belief that the subject acquires in z and her linguistic behavior. It only matters what p means at the actual world.

The closest one can get to simulate the subject's entangled beliefs in the situation t with the disquotational acceptance principle is to represent the uncertainty that the subject has about the semantic facts as an indeterminacy in the meaning of p according to the semantic facts at the actual world. A supervaluation that represents this indeterminacy is the supervaluation $\mathcal{I} = \{I, I'\}$, where $I(p) = \{w_0, w_1\}$ and $I'(p) = \{v_0, v_1\}$. According to $\mathcal{I}$ the meaning of p is indeterminate between expressing the proposition that it is raining and expressing the proposition that it is not raining. This indeterminacy about the meaning of p is analogous to the uncertainty that the subject has about the meaning of p in the metasemantic model above. But the indeterminacy about the meaning of p is not entangled with her beliefs about the weather. Changes to the belief set of the subject do not change the supervaluation representing the meaning of sentences. If we now consider the situation z, where the subject has the belief set $\{w_1\}$, then according to the supervaluation $\mathcal{I}$ the subject does not accepts the sentence p. But in the example we imagine that in z the subject accepts the sentence p because she has learned that it means that it is raining.

The supervaluation that most plausibly represents the language of the subject in the situation z is the singleton set $\{I\}$, where I is as above, that is, $I(p) = \{w_0, w_1\}$. According to this supervaluation it is determinate that p means that it is raining and hence the subject would accept the sentence p in the situation z. Representing the language of the subject in the situation z with this supervaluation $\{I\}$ would make explicit that her language changes from the situation t to the situation z. In z the meaning of p is that it is raining, whereas in t the meaning p was still indeterminate. To model the example in this way would require that we allow the language of the subject to change across different situation. In the notation of families of supervaluations suggested on page 86 in the previous section we would need a family of supervaluations $(\mathcal{I}_w)_{w \in W}$ such that $\mathcal{I}_{w_t} = \{I, I'\}$ and $\mathcal{I}_{w_z} = \{I\}$, where w_t is the actual world of the situation t and w_z is the actual world of the situation z. This account of the example suggests to develop a general theory of how the meaning of sentences in the language of the subject changes. One would need to explain how a change to the belief set of the subject can trigger a change in the supervaluation that represents her language.

The discussion of the above example might give the impression that the ability to represent entanglement in such a way that it influences linguistic behavior should make a difference in the class of behaviors that are interpretable with metasemantic acceptance as opposed to those that are interpretable with disquotational acceptance. However, we know from Theorem 7.5.1 that the class of splitting interpretable behaviors is precisely the same as the class of supervaluation interpretable behaviors. To see how this is possible it is helpful to figure out what kind of supervaluation would interpret the behavior that is metasemantically generated by the metasemantic model discussed in the example.

A supervaluation that interprets the behavior of the subject in the previous example is the supervaluation $\mathcal{I} = \{I\}$ that contains only one interpretation function I such that $I(p) = \{w_1, v_1\}$. To see that this supervaluation generates the same behavior as the metasemantic model above it is helpful to observe that $I(p)$ is precisely the diagonal proposition of the family of interpretation functions in the metasemantic model. According to this interpreting supervaluation the meaning of p is the proposition that the other person accepts the sentence p. Intuitively, this is strange because we would not explain the behavior of the child learning the meaning of p by claiming that she knows already in advance that p means that the person that she is learning from accepts p. But intuitive plausibility of the formal models is not one of the requirements on an account of interpretation from Section 1.4. So we are left looking for an extension of the current setting in which the ability to represent entanglements between beliefs about semantic facts and beliefs about basic facts makes a difference for the class of interpretable behaviors.

In the doxastically richer setting of plausibility orders from Section 3.3 entanglement does lead to a difference between splitting interpretability and supervaluation interpretability. I discuss this matters in Example 7.5.4. There I describe a linguistic behavior that is splitting plausibility interpretable but not supervaluation plausibility interpretable for a suitable, very weak notion of supervaluation plausibility interpretability. This linguistic behavior serves as a mathematical counterexample but I can not find a story that makes it plausible that some subject would show this behavior. Hence it remains an open question whether the ability to represent entanglement is an advantage for an account of interpretation.

References to the literature

The entanglement of semantic and basic facts discussed in this section is analogous to the Frege-Geach problem for expressivist theories in metaethics, which results from a similar entanglement of normative and descriptive facts (see Schroeder 2008 for an overview). The representation of entanglement in metasemantic possible world models is analogous to the solution of the Frege-Geach problem given by Gibbard (2003, ch. 3), who introduces fact-plan worlds that combine a complete specification of all the descriptive facts with a complete specification of all the normative facts.

Chapter 6

Necessity

In this chapter I consider the problem of interpretation when we have a hypothesis about which constructions in the language of the subject function as the necessity modality. I start in Section 6.1 with introducing the problem of interpreting a necessity modality. In Section 6.2 it is shown that if we interpret the necessity modality with a simple account of meaning that follows naturally from the disquotational acceptance principle then we run into a well-know problem with necessity a posteriori. In Section 6.3 we see that no such problem arises if one uses the metasemantic acceptance principle. By introducing a two-dimensional notion of meaning one can also avoid the problem of necessity a priori for disquotational acceptance. Section 6.4 contains such an account of interpretation based on two-dimensional meanings. Lastly, in Section 6.5 I evaluate and compare the accounts of interpretation from Sections 6.3 and 6.4 with respect to the requirements from Section 1.4.

6.1 The necessity modality

In this section I explain the formal language of modal logic and the intended semantics of the modal necessity operator. These are needed for the discussion in the following sections where we interpret the linguistic behavior of some subject for which we have a hypothesis about which expression in her language correspond to the modal necessity modality.

In this chapter I use the necessity modality as an example of a modality in the language of the subject. I focus on the necessity modality and not for instance on belief ascriptions, epistemic modals or deontic modalities because it has a simple formal semantics. Especially, I need that the necessity modality embeds nicely into complex sentences and that it does not give rise to metasemantic readings.

When we are interpreting the linguistic behavior of the subject with a hypothesis about what expressions in her language correspond to a modal necessity operator this amounts to assuming that the sentences in the vocabulary of the subject have the form of modal formulas that contain one unary modal operator standing for necessity. This assumption is analogous to the treatment of propositional connectives in the previous sections of this thesis.

The modal formulas that I use here are built from atomic sentences in the set At using the propositional connectives, which we discuss in Section 1.2, and one unary modal operator $\Box$. This allows us to use formulas such as $\Box p$ and $\neg\Box(q \to p)$, which might for instance correspond to the English sentences "It is necessary that it is raining." and "It is not necessary that if there are raindrops on the window then it is raining. The set of all modal formulas that can be formed from the set of atomic sentences At using the propositional connectives and the $\Box$-operator is written as $\mathcal{M}$. If we have a hypothesis about what expression in the language of the subject are the propositional connectives and the modal operator then we are working in the case where $\mathcal{V} = \mathcal{M}$.

The intended semantic clause for the necessity operator relative to an interpretation function $I : \mathcal{M} \to \mathcal{P}W$ is as follows:

$$I(\Box\varphi) = \left\{ \begin{array}{ll} W, & \text{if } I(\varphi) = W, \\ \emptyset, & \text{otherwise,} \end{array} \right\} = \{w \in W \mid v \in I(\varphi) \text{ for all } v \in W\}.$$

This clause requires that the proposition expressed by $\Box\varphi$ is true at some world precisely if the proposition expressed by φ is true at all worlds in the domain W.

We can use this semantic clause for the necessity modality together with the semantic clauses for the propositional connectives from Section 1.2 to extend an interpretation $I : \mathsf{At} \to \mathcal{P}W$ that is defined just for the atomic sentences to an interpretation $I' : \mathcal{M} \to \mathcal{P}W$ that satisfies all the semantic clauses. Hence, similar to the situation in Section 1.2, we can work with interpretation functions $I : \mathsf{At} \to \mathcal{P}W$ and assume them to be functions $I : \mathcal{M} \to \mathcal{P}W$.

If we have a hypothesis about what expressions in the language of the subject correspond to the necessity modality then we can assume that the linguistic behavior that we are interpreting is a function $a : S \to \mathcal{P}\mathcal{M}$ that maps a situation $s \in S$ to the set $a(s) \subseteq \mathcal{M}$ of all the modal formulas that correspond to a sentences that the subject accepts in s. In the following two sections of this chapter I explain how to combine the above semantic clause with either disquotational or with metasemantic acceptance to obtain an account of interpretation for behaviors that contain the necessity modality.

References to the literature

The semantic clause for the necessity modality from this section, and the more refined semantics that are given in the following two sections, are commonly taking to account for a special notion of necessity called metaphysical necessity. Starting with Putnam (1975) and Kripke (1980) metaphysical necessity has been distinguished from other notions of necessity such as epistemic, analytic or logical necessity. It has however proven quite difficult to establish what metaphysical necessity is.

In this thesis I do not commit to any substantial view about the nature of metaphysical necessity. From the perspective of developing an account of interpretation finding the right notion of metaphysical necessity simply amounts to finding a formal semantics that allows us to interpret the necessity modality in the language of the subject. This is similar to the approach of Rayo (2013, sec. 2.2.1) that relates metaphysical necessity to the acceptance of 'just is'-statements in English. The general idea is that in accepting or rejecting 'just is'-statements such as for instance "For Susan to be a sibling just is for her to share a parent with someone else." or "For a glass to be filled with water just is for it to be filled with H_2O." we delineate the logical space of cases that we consider in theoretical investigations. The acceptance of the latter 'just is'-statement for example excludes the case where there is water but not H_2O in a glass as a relevant possibility for theoretical inquiry. Rayo suggests to relate metaphysical necessity to the acceptance of 'just is'-statements by postulating that a sentence is metaphysically possible if and only if it is consistent with all accepted 'just is'-statements. Hence by duality a sentence is metaphysically necessary if it is a consequence of the accepted 'just is'-statements.

6.2 Necessity a posteriori

This section gives a straight-forward account of interpretation for the necessity modality based on the disquotational acceptance principle. It turns out that this account suffers from a serious defect with respect to the variety requirement because it presupposes that the subject accepts the same necessity statements no matter what she believes about the basic facts.

The account of this section combines supervaluation interpretability from Section 5.2 with the semantic clause for the necessity operator from the previous section. We are interpreting with supervaluation models as defined in Definition 5.2.2 and use Definition 5.2.3 for the notion of the behavior generated by some supervaluation model. In this section we are interested in behaviors $a : S \to \mathcal{PM}$ and so we need the semantic clause discussed in the previous section to evaluate formulas that contain the $\Box$-operator relative to an interpretation function that is given only on the atomic sentences in At. We can then consider supervaluation interpretability of some behavior with a belief function as defined in Definition 5.2.4.

I do not have a representation results that characterizes the class of behaviors $a : S \to \mathcal{PM}$ that are supervaluation interpretable. In Proposition 7.6.1 it is however proven that supervaluation interpretability entails a simple condition on behaviors $a : S \to \mathcal{PM}$ that is too strong. It is as follows:

6.2.1. Definition. A behavior $a : S \to \mathcal{PM}$ satisfies *that necessity is a priori* if for all situations $s, t \in S$ and sentences $\varphi \in \mathcal{M}$:

$$\text{If } a(s) \text{ is consistent and } \Box\varphi \in a(s) \text{ then } \Box\varphi \in a(t).$$

This condition requires that in all situation where the subject is consistent she accepts the same sentences as being necessarily true.

There are examples showing that it is implausible that necessity is a priori. It follows that the account suggested here does not satisfy the variety requirement. Let me give one such example. We imagine that the subject is a chemist in the 18th century when it was discovered that the molecular structure of water is H_2O. The first situation s_1 is shortly before the discovery that water is H_2O and so the subject still considers it possible that water might have a different molecular structure, say for instance XYZ. In the second situation s_2 the molecular structure of water has been discovered and so the subject knows that water is H_2O. The subject is a speaker of English. Hence her language contains the sentence "Water is H_2O." for which we use the letter p. The subject has the metaphysical view that water necessarily has the molecular structure that it actually has. Thus she accepts the sentence $\Box p$, which corresponds to the English "It is necessary that water is H_2O.", in situation s_2 because there she knows that water is H_2O. Moreover, we can assume that the set of sentences that she accepts in s_2 is consistent. If necessity was a priori it should follow that the subject also accepts $\Box p$ in s_1. But in s_1 the subject does not even believe that water is H_2O and hence she does not accept the sentence $\Box p$ which is the English "It is necessary that water is H_2O."

The example from the previous paragraph demonstrates that the most straightforward account of interpretation that combines disquotational acceptance with the semantic clause for necessity from Section 6.1 fails on the variety requirement because it requires that necessity is a priori. In Section 6.4 I discuss an improvement of the account that uses a two-dimensional semantics for the necessity modality to get rid of this condition. Before I however first show in Section 6.3 that a natural account of interpretation that combines metasemantic acceptance with the semantic clause from Section 6.1 does not lead to the difficulties with necessity a priori.

References to the literature

The example used above to argue against the requirement that necessity is a priori is based on the examples given by Putnam (1975). Similar examples have also been discussed by Kripke (1980). A possible conclusion from such examples is that the necessity modality does not quantify over all the possible worlds that are needed to account for the uncertainty of the subject. This conclusion might be formulated as the claim that metaphysical necessity is a more restrictive than epistemic necessity and it has been defended for instance Putnam (1975, p. 151), Burge (1986, sec. 1) or Soames (2005, p. 329).

Let me sketch how we could turn the idea that metaphysical necessity is more restrictive than epistemic necessity into an account of interpretation that does not require necessity to be a priori. For this we partition the set of possible worlds into equivalence classes that restrict the quantification of the necessity modality. An interpreting supervaluation model would then be of the form $(W, R, b, \mathcal{I})$, where $(W, b, \mathcal{I})$ is a standard supervaluation model and $R \subseteq W \times W$ is an equivalence relation on W. The semantics of the necessity modality is changed such that it quantifies only over the equivalence class of the current world instead of quantifying over all worlds. This yields the following semantic clause:

$$I(\Box\varphi) = \{w \in W \mid v \in I(\varphi) \text{ for all } v \in W \text{ with } wRv\}.$$

The resulting account of interpretation no longer requires that necessity is a priori because different doxastic alternatives of the subject can be in different equivalence classes of the relation R and hence make different formulas of the form $\Box\varphi$ true.

There are two reasons why I do not develop this account in this thesis. First, in Section 6.4 I present a different account for interpreting a necessity modality that maintains disquotational acceptance but does not enforce that necessity is a priori. It uses a two-dimensional semantics and hence is more easily comparable to the account of Section 6.3. The second reason for not developing the account which restricts the quantification of the necessity modality is that I have not been able to find a concise characterization of interpretability in this setting. The difficulty is that the account seems to impose cardinality constraints that lead to an intricate interaction between the acceptance of sentences in the scope of the necessity modality and the acceptance of these sentence across different situations. I explain this rather technical point more extensively at the end of Section 7.6.

There is yet another possibility for dealing with necessity a posteriori if one uses disquotational acceptance. If we would allow that the meanings of sentences change across different situations then it would no longer follow that necessity is a priori. For instance in the example from this section we could say that the meaning of "water" changes with the discovery that water is H_2O. Before the discovery that water is H_2O the term water still applies to substances with molecular structure other than H_2O and hence the subject does not accept "It is necessary that water is H_2O." After the discovery that water is H_2O the meaning of "water" has changed such that it only applies to H_2O and hence the subject accepts "It is necessary that water is H_2O." Putnam (1975, p. 142) explicitly rejects such an account of his example. Nevertheless it would be interesting to see whether this idea can be developed further to obtain a general theory of when and how meanings in the to language of the subject change.

6.3 No problem for metasemantic acceptance

In this section I combine the semantic clause for the necessity modality from Section 6.1 with the metasemantic acceptance principle. The resulting account of interpretation does not face the difficulties with necessity a posteriori that affects the account from the previous section.

To use metasemantic acceptance in an account of interpretation for behaviors that contain modal sentences we need to use metasemantic possible world models. It is no longer sufficient to use the single interpretation of a multi-situation possible world model to encode the diagonal proposition of sentences as described in the end of Section 4.3. With the semantic clause for necessity from Section 6.1 the truth value of a sentences of form $\Box\varphi$ at some world depends on the whole proposition expressed by the sentence φ at that world. It is not sufficient to know just the truth value that the sentence φ has at some world to determine the truth value of $\Box\varphi$ at that world.

Let me give an abstract example to illustrate this point. Take a pair of worlds $W = \{w, v\}$ as the domain and consider just one sentence p. Let the semantic facts obtaining at these worlds be represented by a family of interpretations such that $I_w(p) = \{w, v\}$ and $I_v(p) = \{v\}$ which is given by the following table:

	w	v
w	p	p
v	$\neg p$	p

With the semantic clause for $\Box p$ it follows that $w \in I_w(\Box p)$ because both w and v are in $I_w(\Box p)$. This depends on it being the case that $v \in I_w(p)$, which concerns an entry that is not on the diagonal of the table. For the other world we have that $v \notin I_v(\Box p)$ because at the non-diagonal entry it holds that $w \notin I_v(p)$. Hence the truth value of $\Box p$ at some worlds depends on the whole proposition that is expressed by p. To determine whether $\Box p$ is true at some world it does not suffice to know the diagonal proposition of p.

The example also shows that for modal formulas the diagonal proposition does not satisfy the semantic clauses for an interpretation function, which holds for propositional formulas as observed on page 71 in Section 4.3. The function $D : \mathcal{M} \to \mathcal{P}W$ that maps a sentence to its diagonal proposition does not satisfy the semantic clause for the necessity modality because we have that $D(p) = \{w, v\}$ but $D(\Box p) = \{w\}$ and hence $v \notin D(\Box p)$.

To give an account of interpretation using metasemantic possible world models we need to adapt the definitions from Chapter 2 that lead to the notion of splitting interpretability for multi-situation models. First we need the metasemantic analogue of a splitting interpretation model from Definition 2.4.2:

6.3.1. Definition. A *splitting family of interpretation functions* over W is a tuple $(W', f, (I_w)_{w \in W'})$, where W' is a splitting of W with the surjective splitting function $f : W' \to W$ and for every $w \in W'$ the function $I_w : \mathsf{At} \to \mathcal{P}W'$ is an interpretation that maps atomic sentences to propositions over the split set of worlds W'.

Next we define the notion of interpretability that is an adaptation of Definition 2.6.2 to metasemantic models. In this definition we make use of Definition 4.3.1, for the case where $\mathcal{V} = \mathcal{M}$, which defines the behavior $a^M : S \to \mathcal{PM}$ that is metasemantically generated by a metasemantic possible world model $M = (W, b, (I_w)_{w \in W})$.

6.3.2. Definition. A splitting family of interpretations $(W', f, (I_w)_{w \in W'})$ *interprets a linguistic behavior* $a : S \to \mathcal{PM}$ *with a belief function* $b : S \to \mathcal{P}W$ if $a = a^M$ for the metasemantic model $M = (W', b', (I_w)_{w \in W'})$, where $b' : S \to \mathcal{P}W'$ is defined such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$.

A linguistic behavior $a : S \to \mathcal{PM}$ is *metasemantically splitting interpretable with a belief function* $b : S \to \mathcal{P}W$ if there is some splitting family of interpretation functions (W', f, I) that interprets a with b.

Let us consider an example that illustrates metasemantic splitting interpretability and that demonstrates that the problematic behavior from the previous section is interpretable. Take a domain $W = \{w, v\}$ that contains two possible worlds. The first w is like our actual world in that there the molecular structure of water is H_2O. At v the molecular structure of water is XYZ. We are only interested in one sentence p for which we define a family of interpretation functions such that $I_w(p) = \{w, v\}$ and $I_v(p) = \emptyset$. The same information is represented by the following table:

	w	v
w	p	p
v	$\neg p$	$\neg p$

One might take the sentence p in this table to stand for the English sentence “Water is H_2O.” This sentence is necessarily true at the actual world w because water is H_2O, but if the molecular composition of water had been different then the sentence would have expressed a proposition that is necessarily false.

We can turn this family of interpretations into a metasemantic possible world model $M = (W, b, (I_w)_{w \in W})$ by adding a belief function $b : S \to \mathcal{P}W$. For this we use the belief function from the counterexample in the previous section where $S = \{s_1, s_2\}$, $b(s_1) = \{w, v\}$ and $b(s_2) = \{w\}$. So at s_1 the subject is uncertain about the molecular composition of water but at s_2 she knows that it is H_2O. Because $I_w(\Box p) = \{w, v\}$ and $I_v(\Box p) = \emptyset$ we have that $D(\Box p) = \{w\}$. With the definition of the behavior a^M metasemantically generated by M it follows that $\Box p \in a^M(s_2)$ but $\Box p \notin a^M(s_1)$. This shows that a^M is the kind of behavior that is discussed as a counterexample to necessity a priori in the end of the previous section. One can see that this behavior is indeed metasemantically splitting interpretable with b in the sense of Definition 6.3.2 by considering the splitting family of interpretation functions $(W, f, (I_w)_{w \in W})$ where $f : W \to W$ is be the identity function such that $f(w) = w$ for all $w \in W$.

The account from this section avoids the condition that necessity is a priori even though it uses the semantic clause from Section 6.1 to interpret the necessity modality. This is possible because with the metasemantic acceptance principle the subject's acceptance of sentences is influenced by her beliefs about the meaning of these sentences. But these beliefs can change depending on the situation in which we are interpreting the subject. Hence it is possible that in different situations the subject accepts different sentences as being necessarily true, even thought the necessity modality always quantifies over the same set of all possible worlds.

Theorem 7.6.9 gives a general characterization of the behaviors that are metasemantically interpretable with some belief function. It shows that a behavior $a : S \to \mathcal{PM}$ is metasemantically splitting interpretable with a belief function $b : S \to \mathcal{P}W$ if and only if it satisfies the conjunctive covering condition and the conjunctive consistency condition that are discussed in Section 2.6. These conditions make use of the notions of logical consequence and consistency. To apply these notions in the context of this chapter, where $\mathcal{V} = \mathcal{M}$, we need to define a notion of logical consequence between the modal formulas in $\mathcal{M}$. The notion of logical consequence between modal formulas that is needed for the representation result in Theorem 7.6.9 is logical consequence in the modal logic S5. This is not surprising since we evaluate the necessity modality with the semantic clause from Section 6.1.

Logical consequence in S5 can be defined completely analogously to the definition of consequence in propositional logic from Section 1.2. A sentence $\varphi \in \mathcal{M}$ is a *consequence in S5* of a set of sentences $\Sigma \subseteq \mathcal{M}$ if for every set of worlds W and interpretation $I : \mathsf{At} \to \mathcal{P}W$ it holds that $\bigcap\{I(\psi) \mid \psi \in \Sigma\} \subseteq I(\varphi)$. What makes this definition different from the definition in the propositional case is that in order to evaluate modal sentences we need to use the semantic clause from Section 6.1. Once a notion logical consequence between modal formulas is defined we obtain a notion of consistency in the obvious way. A set $\Sigma \subseteq \mathcal{M}$ is *consistent in S5* if $\bot$ is not a consequence in S5 of Σ.

The representation theorem especially shows that there are no constraints for metasemantic splitting interpretability that require necessity to be a priori. This suggests that the account from this section satisfies the variety requirement. In Section 6.5 I give a more extensive evaluation of metasemantic splitting interpretability on the requirements from Section 1.4.

References to the literature

That one can use a setting similar to metasemantic possible world models to account for necessity a posteriori is a key insight of Stalnaker (1978). This section reformulates Stalnaker's ideas in the context of a theory of interpretation.

6.4 Two-dimensional meanings

In this section I present an account of interpretation for linguistic behaviors containing a necessity modality that is based on the disquotational acceptance principle but does not require that necessity is a priori. Formally, the resulting account is similar to the families of interpretation functions that are employed in the notion of metasemantic splitting interpretability discussed in the previous section. But on a conceptual level one thinks differently about the formal models.

The account of this section uses are more complex notion of meaning than just sets of worlds. It distinguishes two kinds of dependency of the truth-value of some sentence on possible worlds. The first is called *epistemic dependency* and it captures the change of the truth value of some sentence with respect to the information that the subject has about the world. The second kind of dependency is called *counterfactual dependency* and it determines how some sentence behaves when it occurs embedded in the scope of a necessity modality.

To capture these two kinds of dependencies we take meanings to be sets of pairs of worlds instead of just sets of worlds. Variance in the first component of this pairs captures the epistemic dependency that sentences have on the information that the subject has about the world, whereas variance in the second component captures the counterfactual dependency that regulates how the sentences embeds in the scope of a necessity modality. We have to change the definition of an interpretation function to accommodate for this more complex notion of meaning:

6.4.1. Definition. A *two-dimensional interpretation function* is a function $I : \mathcal{V} \to \mathcal{P}(W \times W)$ that maps sentences to sets of pairs of worlds.

Intuitively, one thinks of a two-dimensional interpretation function for the language of the subject to be such that $(u, w) \in I(\varphi)$ if and only if in a situation where the subject believes that u is the actual world she thinks that if w had been the case then φ would be true. More concisely, we might say that $(u, w) \in I(\varphi)$ if and only if from the perspective of u the sentence φ is true at w.

By defining semantic clauses one can extend a two-dimensional interpretation $I : \mathsf{At} \to \mathcal{P}(W \times W)$ that is defined just for atomic sentences to a two-dimensional interpretation $I : \mathcal{M} \to \mathcal{P}(W \times W)$ that is defined on all modal formulas over the set of atomic sentences. The semantic clauses for the propositional connectives are analogous to the clauses for standard interpretation functions from Section 1.2. For instance we require that $I(\varphi \wedge \psi) = I(\varphi) \cap I(\psi)$ or that $I(\neg\varphi) = (W \times W) \setminus I(\varphi)$. The semantic clause for the necessity modality universally quantifies over the second of the two components while keeping the first component fixed:

$$I(\Box\varphi) = \{(u, w) \in W \times W \mid (u, v) \in I(\varphi) \text{ for all } v \in W\}.$$

This semantic clause captures the idea that only the variance in the first components of pairs of worlds influences how a sentence embeds into the scope of the necessity modality.

Let us consider an example of a two-dimensional interpretation function. We represent the two-dimensional meaning of the sentence p that we think of the English sentence "Water is H_2O." The domain $W = \{w, v\}$ contains two possible worlds. The world w is like our actual world in that there water is H_2O. At v the molecular structure of water is XYZ. To determine the two-dimensional meaning of p we have to ask ourselves for each pair of possible worlds whether we would accept that p is true at the second component of the pair if we had the knowledge that the first component of the pair is the actual world. First consider the pair (w, w). We need to ask us whether we would say that "Water is H_2O." is true at w if we know that water has molecular structure H_2O. Clearly this is the case and hence (w, w) is in the two-dimensional meaning of p. For the pair (v, w) we ask ourselves whether under the assumption that we know that water has molecular structure H_2O we would accept that "Water is H_2O." correctly describes the world v. Let us assume that this is the case and hence (v, w) is in the two-dimensional meaning of p. If on the other hand we suppose that we knew that the molecular structure of water XYZ then we would judge the sentence "Water is H_2O." to be false both at the world where water is H_2O and at the world where it is XYZ. Hence neither (w, v) and (v, v) is in the two-dimensional meaning associate to p. Summarizing this paragraph we have that $I(p) = \{(w, w), (v, w)\}$.

From the semantic clause for $\Box p$ relative to I it follows that $\Box p$ is true precisely at those pairs that have w as the second component. Whenever we know that water has molecular structure H_2O we would also accept that the sentence "It is necessary that water is H_2O."

One can encode two-dimensional interpretation functions with the help of tables. The rows in such a table correspond to variance of truth-value in the first component of the pairs in its two-dimensional meaning and the columns represent the variance in the second component. This means that the truth value of some sentence at the pair $(u_1, u_2) \in W \times W$ is written in the cell of the row indexed by u_1 and the column indexed by u_2.

As an example we obtain the following table for the meaning of p in the two-dimensional interpretation I from the example above:

	w	v
w	p	p
v	$\neg p$	$\neg p$

This table is the same as the table encoding the family of interpretation functions for an analogous example in Section 6.3. This suggests that there is a close connection between two-dimensional interpretations and families of interpretations. Formally this is indeed the case and is the reason why, as we see in the following section, the two settings give rise to similar accounts of interpretation. Despite these formal similarities there is a crucial conceptual difference between the settings. A two-dimensional interpretation represents the meaning of sentences according to the semantic facts at the actual world. This meanings are complex two-dimensional structures that encode both counterfactual and epistemic dependencies of truth values. A family of interpretations on the other hand represents the meanings of sentences relative to all the worlds in the domain. The meanings themselves are simple propositions that only encode the counterfactual dependency of truth-values of sentences.

There are still two obstacles to using two-dimensional interpretations in an account of interpretation.

The first obstacle is that it is unclear when the subject believes the two-dimensional meanings that are associated to sentences in her language. In the setting of this thesis the subject believes propositions which are sets of worlds instead of sets of pairs of worlds. We need to transform sets of pairs of worlds into a sets of worlds. The canonical way of doing so is to consider the diagonal proposition in a set of pairs of worlds. This is analogous to the diagonal proposition in a family of interpretation functions and can be defined as follows:

6.4.2. Definition. The *diagonal proposition* $X^d \subseteq W$ *of an* $X \subseteq W \times W$ is defined such that

$$X^d = \{w \in W \mid (w, w) \in X\}.$$

The diagonal proposition is sensitive to the epistemic dependency in the first component of the pairs that are in the meaning of some sentence. In the example above we have that the diagonal proposition of "Water is H_2O." is the set $(I(p))^d = \{w\}$ which only contains the world at which water has the molecular structure H_2O and excludes the world where water is XYZ. Nevertheless, we still have that $(I(\Box p))^d = \{w\}$ because the semantics of the necessity modality is sensitive to the counterfactual variance in the second component and not to the epistemic variance in the first component.

We model the proposition expressed by a sentence $\varphi \in \mathcal{M}$ relative to a two-dimensional interpretation function $I : \mathsf{At} \to \mathcal{P}(W \times W)$ as the diagonal proposition $(I(\varphi))^d \subseteq W$. If we apply the acceptance principle it follows that a subject with belief function $b : S \to \mathcal{P}W$ accepts some sentence φ in some situation s if and only if $b(s) \subseteq (I(\varphi))^d$.

The second obstacle to using two-dimensional interpretations in an account of interpretation is that we need to avoid the cardinality condition on tight interpretability to cope with the problem of vagueness. As a solution I adapt the supervaluations from Section 5.2 to the setting of two-dimensional interpretations. The reason for using supervaluations and not some notion of splitting interpretability is that, as explained in Sections 5.1 and 5.2, supervaluations fit better into a setting that uses disquotational acceptance.

We change the notion of a supervaluation from Definition 5.2.1 to contain two-dimensional interpretation functions instead of standard interpretation functions.

6.4.3. Definition. A *two-dimensional supervaluation $\mathcal{I}$ over W* is a non-empty set of two-dimensional interpretation functions such that every $I \in \mathcal{I}$ is a function $I : \mathsf{At} \to \mathcal{P}(W \times W)$.

We also adapt the notion of a supervaluation model from Definition 5.2.2 to contain two-dimensional supervaluations:

6.4.4. Definition. A *two-dimensional supervaluation model* $M = (W, b, \mathcal{I})$ is a domain W together with a belief function $b : S \to \mathcal{P}W$ and a two-dimensional supervaluation $\mathcal{I}$ over W.

The definition of the behavior generated by some two-dimensional supervaluation model is based on the disquotational acceptance principle for supervaluations from Section 5.2. In this principle we have to understand the proposition expressed by a sentence to be the diagonal proposition as defined above. Thus we obtain the following variant of Definition 5.2.3:

6.4.5. Definition. The *linguistic behavior $a^M : S \to \mathcal{PM}$ generated by a two-dimensional supervaluation model* $M = (W, b, \mathcal{I})$ is defined such that for all $s \in S$

$$a^M(s) = \{\varphi \in \mathcal{V} \mid b(s) \subseteq (I(\varphi))^d \text{ for all interpretations } I \in \mathcal{I}\}.$$

Finally, we define a notion of interpretability that is analogous to supervaluation interpretability from Section 5.2.

6.4.6. Definition. A two-dimensional supervaluation $\mathcal{I}$ over W *interprets a linguistic behavior $a : S \to \mathcal{PM}$ with a belief function $b : S \to \mathcal{P}W$* if $a = a^M$ for the two-dimensional supervaluation model $M = (W, b, \mathcal{I})$.

A linguistic behavior $a : S \to \mathcal{PV}$ is *two-dimensional supervaluation interpretable with a belief function $b : S \to \mathcal{P}W$* if there exists some two-dimensional supervaluation that interprets a with b.

As an example we show that the behavior from Section 6.2, which is a counterexample to necessity a priori, is two-dimensional supervaluation interpretable. Let p stand for the sentence "Water is H_2O." and consider the domain $W = \{w, v\}$ such that at w the molecular structure of water is H_2O and at v it is XYZ. We consider the subject in two situations s_1 and s_2. At s_1 she is uncertain about the molecular structure of water and so her belief function $b : S \to \mathcal{P}W$ is such that $b(s_1) = \{w, v\}$. At s_1 she knows that the chemical composition of water is H_2O and hence $b(s_2) = \{w\}$.

We interpret the subject with the two-dimensional supervaluation $\mathcal{I} = \{I\}$ that contains only the two-dimensional interpretation $I : \mathsf{At} \to \mathcal{P}(W \times W)$ from the example above. Hence I is defined such that $I(p) = \{(w, w), (w, v)\}$. Now consider the two-dimensional supervaluation model $M = (W, b, \mathcal{I})$. The behavior $a^M : S \to \mathcal{PM}$ generated by this model is such that $\Box p \in a^M(s_2)$ but $\Box p \notin$

$a^M(s_1)$. Hence the account from this section can account for behaviors that do not satisfy that necessity is a priori.

The representation result for two-dimensional supervaluation interpretability is given by Theorem 7.6.13. It states that under the assumption that At is at most countably infinite a linguistic behavior $a : S \to \mathcal{PM}$ is two-dimensional supervaluation interpretable with a belief function $b : S \to \mathcal{P}W$ if and only if it satisfies the conjunctive covering and consistency conditions, where the notion of logical consequence that is used in these conditions is consequence in the modal logic S5 in the case where W is infinite and it is consequence in the modal logic S5.n in the case where W is a finite set with n elements.

Syntactically, the modal logic S5.n is an extension of the logic S5. I do not give a syntactic definition of the logic S5.n in this chapter. The interested reader is referred to Definition 7.6.10. The only thing that I want to mention here is that there is a close similarity between the additional axioms of S5.n, which are not in S5, and the cardinality condition on tight interpretability from Chapter 2. The size of W is an upper bound on how many counterfactual worlds the subject can think there exist. For two-dimensional supervaluation interpretability the subject can not by her acceptance of modal sentences imply that there are more possible world than there are in the domain W.

As a semantic definition of logical consequence in S5.n we can constrain the definition of consequence in S5 from the previous section. A sentence $\varphi \in \mathcal{M}$ is a *consequence in S5.n* of a set of sentences $\Sigma \subseteq \mathcal{M}$ if for every finite set W containing at most n worlds and interpretation $I : \mathsf{At} \to \mathcal{P}W$ it holds that $\bigcap\{I(\psi) \mid \psi \in \Sigma\} \subseteq I(\varphi)$. Here we are interpreting the modality with the semantic clause from Section 6.1. Alternatively, one could also quantify over all domains W containing at most n worlds and two-dimensional interpretations $I : \mathsf{At} \to \mathcal{P}(W \times W)$ and then use the semantic clause from this section for the modality on two-dimensional interpretations.

In the following section I discuss how the account from the section performs on the requirements from Section 1.4 and compare it to the account from the previous section.

References to the literature

The two-dimensional conception of meaning that is employed in this section comes from the epistemic interpretation of two-dimensional semantics that is developed for instance by Chalmers (2002; 2006). A central claim of epistemic two-dimensionalists is that the epistemic significance of a sentence is captured by the diagonal proposition in the two-dimensional meaning associate to the sentences. This is accounted for in the setting of this because we compare the diagonal proportion of some sentence with the belief set of the subject to determine whether the subject accepts the sentence.

Chalmers (2006, sec. 5.1) compares his understanding of the two-dimensional framework to the one of Stalnaker that is the basis for the metasemantic models from Section 4.1 and for the notion of metasemantic splitting interpretability from the previous section. A similar comparison, from the other side, is given by Stalnaker in (2001, secs. 3–5) and (2004, secs. 3 and 4). Chalmers and Stalnaker are mainly concerned with conceptual differences in their understanding of the two-dimensional framework. My comparison in the following section focuses more on how the two approaches perform as accounts of interpretation.

6.5 Evaluation and comparison

In this section I evaluate the accounts of interpretation from the previous two sections on the requirements from Section 1.4. Because of the formal similarity of splitting families of interpretation functions and two-dimensional supervaluations there are little differences between the two accounts.

I start with the little-input requirement. Both, metasemantic splitting interpretability and two-dimensional supervaluation interpretability, suffer from two problems with the little-input requirement that I want to mention here, even though I do not solve them in this thesis.

The first is that we assuming to know in advance to interpretation which constructions in the language of the subject function as the necessity modality. It would be better if we had a mechanism that parses sentences in the language of the subject and finds the modalities as part of the process of interpretation. This is analogous to the difficulty with finding the propositional connectives in the language of the subject that is already mentioned in Section 2.1.

The other problem is that we assume to know all beliefs of the subject for every situation where we are interpreting her. This assumption is also made in the account from Section 2.6, and one might try to weaken the assumption similarly as in Chapter 3 by introducing evidence functions and plausibility orders.

Also with respect to the determinacy requirement metasemantic splitting interpretability and two-dimensional supervaluation interpretability give rise to similar difficulties. Both accounts leave room for an indeterminacy that might be undesirable. The problem is roughly that the acceptance of modal sentences does not determine how the counterfactual worlds as described in the language of the subject match up with the worlds in the original domain relative to which the basic facts are fixed. Let me explain this indeterminacy more carefully in the case of metasemantic splitting interpretability and then show how it affects two-dimensional supervaluation interpretability.

In the setting of families of interpretation functions the indeterminacy is as follows: Consider any world w in some domain W that comes with a family of interpretation functions $(I_w)_{w \in W}$ representing the semantic fact at worlds in W. We can change the interpretation function I_w representing the semantic facts at w without affecting the linguistic behavior generated by a metasemantic model that is based on $(I_w)_{w \in W}$. To do so take any two worlds v and u that are distinct from each other and distinct from w. Now consider a new family of interpretation functions $(I'_w)_{w \in W}$ which is defined such that $I'_{w'} = I'_{w'}$ for every w' that is distinct from w and I_w is such that for all atomic sentences p it holds that $v \in I'_w(p)$ if and only if $u \in I_w(p)$, $u \in I'_w(p)$ if and only if $v \in I_w(p)$ and $w' \in I'_w(p)$ if and only if $w' \in I_w(p)$ for every w' that is distinct from u and v. The idea behind the definition is to let u and v completely switch their roles with respect to the sentences they make true according to the semantic facts at w. We obtain an indeterminacy because that the linguistic behavior metasemantically generated by some metasemantic model $M = (W, b, (I_w)_{w \in W})$ that based on $(I_w)_{w \in W}$ is the same as the behavior generated by the model $(W, b, (I'_w)_{w \in W})$ that is just like M with the only difference that $(I_w)_{w \in W}$ is replaced with $(I'_w)_{w \in W}$

For an example consider again the domain $W = \{w, v, u\}$ from the first part of this thesis, where w is a world at which it is raining and there are raindrops on the window, at v it is not raining but there are still raindrops on the window and at u is neither raining nor are there raindrops on the window. Let p stand for the sentence "It is raining." and q for the sentence "There are raindrops on the window." Assume that at all worlds the semantic facts are as in English. This determines the family $(I_w)_{w \in W}$ of interpretations such that $I_w(p) = I_v(p) = I_u(p) = \{w\}$ and $I_w(q) = I_v(q) = I_u(q) = \{w, v\}$. The table corresponding to $(I_w)_{w \in W}$ is as follows:

	w	v	u
w	$p,\ q$	$\neg p,\ q$	$\neg p, \neg q$
v	$p,\ q$	$\neg p,\ q$	$\neg p, \neg q$
u	$p,\ q$	$\neg p,\ q$	$\neg p, \neg q$

A speaker of English that has certain beliefs about the rain and raindrops on the window could be interpreted with this family of interpretation functions, where we might use the identity function as the splitting because in English there is no uncertainty about the truth values of p and q relative the worlds in W.

We now modify this family to obtain a new family $(I'_w)_{w \in W}$ of interpretation functions such that the semantic facts at w and u stay the same, that is, $I'_w = I_w$ and $I'_u = I_u$, but for v we define that $I'_v(p) = \{u\}$ and $I'_v(q) = \{v, u\}$. This modified family $(I'_w)_{w \in W}$ corresponds to the table:

	w	v	u
w	$p,\ q$	$\neg p,\ q$	$\neg p, \neg q$
v	$\neg p, \neg q$	$\neg p,\ q$	$p,\ q$
u	$p,\ q$	$\neg p,\ q$	$\neg p, \neg q$

The table for $(I'_w)_{w \in W}$ differs from the table for $(I'_w)_{w \in W}$ in that in the row belonging to v the roles of the non-diagonal entries for w and u are swapped.

One can check that for every $\varphi \in \mathcal{M}$ and $v \in W$ we have that $v \in I'_v(\varphi)$ iff $v \in I_v(\varphi)$. It follows that for any belief function $b : S \to \mathcal{P}W$ the behavior metasemantically generated by a metasemantic model $(W, b, (I_w)_{w \in W})$ is the same as the behavior metasemantically generated by the metasemantic model $(W, b, (I'_w)_{w \in W})$.

Intuitively, there is a difference between and $(I'_w)_{w \in W}$. According to the latter family of interpretations the semantic facts are such that at v the sentence p expresses the proposition that there are no raindrops on the window and q expresses the proposition that it is not raining. These meanings seem to be different from the meanings that p and q have in English, which is represented by $(I_w)_{w \in W}$.

Note that the indeterminacy observed here is strictly weaker than the indeterminacy from Section 4.3 that also involves metasemantic models with metasemantic acceptance but concerns behaviors that do not contain the necessity modality. There we had that one could arbitrarily change the truth value of sentences at non-diagonal entries of the table without affecting the linguistic behavior that is generated. This is no longer possible in the setting of this section which includes modal sentences. If we would for instance define $I'_v(q) = W$ then the subject would suddenly start to accept the sentence $\Box q$ in a situation where she has the belief set $\{v\}$.

There is an analogous indeterminacy for two-dimensional supervaluations. This is a consequence of the structural similarities between families of interpretation functions and two-dimensional interpretation functions. To describe the indeterminacy formally consider a two-dimensional supervaluation $\mathcal{I}$ and fix a two-dimensional interpretation I in $\mathcal{I}$. Then take three distinct worlds w, v and u and define a new two-dimensional interpretation I' such that for all atomic sentences p we have that $(w, v) \in I'(p)$ if and only if $(w, u) \in I(p)$, $(w, u) \in I'(p)$ if and only if $(w, v) \in I(p)$ and $(w', w'') \in I'(p)$ if and only if $(w', w'') \in I(p)$ for all w' and w'' such that if $w' = w$ then w'' is distinct from both u and v. The idea is that from the epistemic perspective of w the worlds u and v swap their role as counterfactual worlds. Now consider the two-dimensional supervaluation $\mathcal{I}' = (\mathcal{I} \setminus \{I\}) \cup \{I'\}$ in which I is replaced with I'. Because one can show that $(I'(\varphi))^d = (I(\varphi))^d$ for all modal formulas $\varphi \in \mathcal{M}$ it follows that the behavior generated by a two-dimensional supervaluation model containing $\mathcal{I}$ is the same as the behavior generated by the model in which $\mathcal{I}$ is replaced with $\mathcal{I}'$.

To obtain an example of this indeterminacy one can adapt the example that is given above for the case of families of interpretation functions. One just has to think of the two tables from above as encoding two-dimensional interpretation functions I and I'. It then turns out that the behavior generated by a model based on the singleton supervaluation $\mathcal{I} = \{I\}$ is the same as the behavior generated by the model in which $\mathcal{I}$ is replaced with $\mathcal{I}' = \{I'\}$.

There are two different possible reactions to the indeterminacy discussed in the paragraphs above. One reaction is to accept the indeterminacy as showing that there are intuitive differences between models that are not reflected in linguistic behavior. On this view the examples given above do not witness a violation of the determinacy requirement. They just demonstrate that our intuitive notion of sameness of models is too fine grained. The other reaction is to take the examples as showing that the account does not fulfill the determinacy requirement. In this case it would be interesting to find a richer notion of linguistic behavior or stronger assumptions about the interpreter's prior knowledge that can account for the intuitive difference between the models discussed above.

Lastly, consider the variety requirement. I first discuss metasemantic splitting interpretability and then contrast it with two-dimensional supervaluation interpretability.

The conditions for metasemantic splitting interpretability are an adaptation of the conditions on splitting interpretability that, as discussed in Section 2.6, impose plausible constraints on linguistic behaviors. The difference is that for metasemantic splitting interpretability the subject's use of the necessity modality needs to satisfy the axioms of the modal logic S5. But S5 is generally considered to be the right logic for metaphysical necessity. Hence, metasemantic splitting interpretability performs reasonably well on the variety requirement.

The conditions for two-dimensional supervaluation interpretability are similar to the conditions for metasemantic splitting interpretability. If the vocabulary of the subject does not contain more than countably many atomic sentences, which is plausible, and the domain W is infinite then two-dimensional supervaluation interpretability with a belief function $b : S \to \mathcal{P}W$ coincides with metasemantic splitting interpretability with b. This entails that subject's acceptance of modal sentences needs to satisfy the axioms of S5. If the domain W is a finite set with n elements then there are additional constraints for two-dimensional supervaluation interpretability with a belief function $b : S \to \mathcal{P}W$. In such cases the behavior of the subject needs to satisfy the conjunctive covering and the conjunctive consistency conditions with respect to logical consequence in the modal logic S5.n. The logic S5.n extends S5 with additional axioms. Hence two-dimensional supervaluation interpretability is more restrictive than metasemantic splitting interpretability.

The additional axioms of the logic S5.n that are required for two-dimensional supervaluation interpretability limit the number of counterfactual worlds that the subject can suppose to exist by the size of the original domain W that is fixed in advance to interpretation. Metasemantic splitting interpretability does not impose any such cardinality constraints on the subject's acceptance of sentences containing the necessity modality. Even when the domain is finite it only requires that the subject's acceptance of sentences follows the axioms of S5. This difference between the two notions of interpretability might be used to decide which one gives the better account of interpretation. If we can find a plausible linguistic behavior for a finite domain W with n elements according to which the subject acceptance of sentences violates the axioms in S5.n that are not axioms of S5 then we would have a reason to prefer metasemantic splitting interpretability. If on the other hand we can give an argument that there are no such behaviors then this would be a reason to prefer two-dimensional supervaluation interpretability. But unfortunately the situation is not that clear-cut and I am not sure which notion of interpretability should be preferred.

Let me explain the difference between the two notions of interpretability on an example. Suppose that $W = \{w\}$ contains only one possible world and take a belief function $b : S \to \mathcal{P}W$ and a situation $s \in S$ such that $b(s) = \{w\}$. Because W contains only one element it is necessary for two-dimensional supervaluation interpretability with b that the set of sentences $a(s)$ that the subject accepts in s is a theory in the modal logic S5.1. This entails, somewhat analogously to the case of vagueness, that for every $\varphi \in \mathcal{M}$ the subject accepts the sentence $\Box\varphi \vee \Box\neg\varphi$. Because of the S5 axiom $\Box\varphi \to \varphi$ it also entails that for any sentence $\varphi \in \mathcal{M}$ the subject accepts the biconditional $\varphi \leftrightarrow \Box\varphi$, and so the distinction between truth and necessary truth collapses. Semantically, this is obvious because we give the subject only one possible world to represent her beliefs about necessity.

For metasemantic splitting interpretability with the belief function $b : S \to \mathcal{P}W$ from above there are no such strong constraints. It is only required that the subject follows the axioms of S5. To obtain a linguistic behavior that is metasemantically splitting interpretable with b but not two-dimensional supervaluation interpretable consider the splitting family of interpretations $(W', f, (I)_{w \in W'})$, where $W' = \{w_1, w_2\}$, $f(w_1) = f(w_2) = w$, $I_{w_1}(p) = I_{w_2}(p) = \{w_1\}$, $I_{w_1}(q) = \{w_1, w_2\}$ and $I_{w_2}(q) = \emptyset$. This family of interpretations is given by the following table:

	w_1	w_2
w_1	$p, \ \ q$	$\neg p, \ \ q$
w_2	$p, \neg q$	$\neg p, \neg q$

By definition this splitting family of interpretations interprets the linguistic behavior $a^M : S \to \mathcal{P}\mathcal{M}$ generated by the metasemantic model $M = (W', b', (I)_{w \in W'})$, where $b' : S \to \mathcal{P}W'$ is such that $b'(s) = f^{-1}[b(s)] = \{w_1, w_2\}$. But this behavior a^M is not two-dimensional supervaluation interpretable with b because $\Box p \vee \Box\neg p \notin a^M(s)$.

It is not clear whether a subject might plausibly show the behavior a^M that is a counterexample to the additional axioms of S5.n. To make sense of the behavior a^M we can try to understand the kind of beliefs that are represented in the model M that is generating this behavior. Especially, consider the beliefs about the sentence p because they cause the subject to not accept $\Box p \vee \Box\neg p$. According to the semantic facts of any of the worlds in the model M the sentence p is true at w_1 but false at w_2. The the subject believes that the sentence p is contingent even though the original domain W contains only one possible world and hence relative to that domain no sentence can be contingent.

To understand why the subject might take p to be contingent even thought there are no contingencies in the basic facts that are represented in W we need to pay attention to the difference between w_1 and w_2. Both worlds w_1 and w_2 map under the splitting function $f : W' \to W$ to the world w in the original domain. Hence, as I argue in Section 5.1, the two worlds are the same with respect to the basic facts represented in the original domain W and they might differ at most with respect to the semantic facts. In particular they differ because according to the semantic facts at w_1 the sentence q is true at w_1 and w_2, whereas according to the semantic facts at w_2 this sentence is false at w_1 and w_2.

Since p is true at w_1, but false at w_2, and the worlds w_1 and w_2 differ only with respect to the semantic facts that determine the meaning of q it must be that the meaning of p is about the semantic facts that determine the meaning of q. The meaning of p according to every world in the model M is the proposition that q is true, or equivalently in W', the proposition that q is necessarily true. The subject thinks that p is contingently true because she believes that the semantic facts that determine the meaning of the sentence q are contingent.

I am undecided whether this explanation of what is happening in the model M is convincing enough to think that some subject might plausibly show the behavior a^M that is metasemantically generated by the model M. But even if we would accept that it is plausible that we would find some subject that has the beliefs represented in the model M it is not clear that this would show that two-dimensional supervaluation interpretability does not satisfy the variety requirement. We might concede that the behavior a^M in the example is not two-dimensional supervaluation interpretable with the belief function $b : S \to \mathcal{P}W$ that is defined relative to the singleton domain $W = \{w\}$. Maybe the problem is not the notion of two-dimensional supervaluation interpretability but our assumption that we should be able to interpret the subject of the example with such a poor domain of possible worlds. Then it would be better to start from a domain such as W', which explicitly includes the semantic facts that the subject consider to be contingent as part of the basic facts. The question whether we accept a^M as a counterexample to two-dimensional splitting interpretability depends on whether we are comfortable with allowing the subject's own language to be sensitive to the semantic facts that are introduced by splitting interpretability.

If one feels sufficiently uncomfortable with allowing the meanings in the language of the subject to be about meanings themselves then one might even take the existence of examples as the one discussed here to be a reason to refuse metasemantic splitting interpretability. For this one would need the converse of the variety requirement which says that for all interpretable behaviors there is a subject that might plausibly show them. The behavior in the example is metasemantically splitting interpretable but it would be implausible that some subject shows the behavior. This line of reasoning would make us reject metasemantic splitting interpretability as it is defined in Section 6.3. It is however not sufficient to refuse metasemantic acceptance or splitting interpretability in general. With a simple modification of the account it is possible to avoid the problematic sensitivity of meanings to differences in the semantic facts. One just has to additionally require in Definition 6.3.2 that the interpreting splitting family of interpretation functions $(W', v, (I_w)_{w \in W'})$ over W is such for every $w \in W$ and $p \in \mathsf{At}$ there is some $P \subseteq W$ such that $I_w(p) = f^{-1}[P]$. I conjecture that the behaviors that are metasemantically splitting interpretable in this modified sense are exactly the same as the behaviors that are two-dimensional supervaluation interpretable.

The discussion from this section shows that metasemantic splitting interpretability and two-dimensional supervaluation interpretability provide similar accounts of interpretation. There are minor differences with respect to the variety requirement, but it is not clear which notion should be preferred because of these differences. Even if we would favor one of the two notions of interpretability it seems likely that the other notion can be adapted such that it interprets the same class of linguistic behaviors as the preferred notion.

Chapter 7

The representation results

This chapter contains the proofs of the results mentioned in the thesis. The chapter is organized such that every sections contains results that make use of similar mathematical constructions. Section 7.2 is an exception in that it gives a summary of the various conditions on linguistic behaviors in the multi-situation setting that are used throughout this thesis.

Whenever a proof in the following sections presupposes some non-trivial results I am going to mention this in the beginning of the section. In general I however presuppose some familiarity with propositional logic that goes beyond the definitions given in Section 1.2. Let me mention some of these notions and results in the following paragraphs. The reader can find proofs of these results in any decent introduction to propositional logic.

In Section 1.2 complete theories are defined to be consistent theories Σ such that $\varphi \in \Sigma$ or $\neg\varphi \in \Sigma$ for every formula $\varphi \in \mathcal{B}$. I am going to use the notation $\mathsf{MC}(\mathcal{B}) \subseteq \mathcal{P}\mathcal{B}$ for the set of all complete theories in the propositional language $\mathcal{B}$. This notation makes sense because complete theories are also called *maximally consistent theories* because one can show that they are maximal elements in the order over all consistent theories that is given by the inclusion $\subseteq$ of sets.

I am also using two fundamental theorems about propositional theories. The first allows us to reduce logical consequence in propositional logic to finite sets of formulas.

7.0.1. Theorem (Compactness). *If $\varphi \in \mathsf{cl}(\Sigma)$ then there exists a finite $\Sigma' \subseteq \Sigma$ such that $\varphi \in \mathsf{cl}(\Sigma')$.*

The other theorem is a basic tool for constructing complete theories.

7.0.2. Theorem (Lindenbaum's Lemma). *Let Σ be a theory and $\varphi \notin \Sigma$. Then there exists a complete theory Γ such that $\Sigma \subseteq \Gamma$ and $\varphi \notin \Gamma$.*

The proofs of these theorems are part of any decent introduction to propositional logic. It should however be mentioned that Lindenbaum's Lemma is usually proven for a syntactic notion of logical consequence which needs the axiom of choice. In the setting of this thesis, where logical consequence is defined semantically, it follows directly from the definition of logical consequence. Choice is nevertheless necessary to prove the compactness of the semantic consequence relation.

7.1 Single situations

This section contains proofs for the results mentioned in Sections 2.1 and 2.2. They have in common that we are interpreting the subject in just one single situation and hence work with behaviors $A \subseteq \mathcal{V}$ as defined in Definition 2.1.1.

We first consider the setting from Section 2.1 in the case where $\mathcal{V} = \mathsf{At}$.

7.1.1. Proposition. *Every behavior $A \subseteq \mathsf{At}$ is interpretable.*

Proof. Assume we are given a behavior $A \subseteq \mathsf{At}$. We have to define a simple possible world model $M = (W, B, I)$ such that $A = A^M$. We can just take the domain and the belief set to be the singleton set $W = B = \{w\}$. The interpretation $I : \mathsf{At} \to \mathcal{P}\{w\}$ is defined such that

$$I(p) = \begin{cases} \{w\}, & \text{if } p \in A, \\ \emptyset, & \text{if } p \notin A. \end{cases}$$

It is clear from the definition of I that $B = \{w\} \subseteq I(p)$ iff $p \in A$, which means that $A = A^M$. □

The case where $\mathcal{V} = \mathcal{B}$ is a little more interesting.

7.1.2. Proposition. *A behavior $A \subseteq \mathcal{B}$ is interpretable iff A is a propositional theory.*

Proof. Fix a behavior $A \subseteq \mathcal{B}$. We need to find a model $M = (W, B, I)$ such that $A = A^M$. The domain of this model is the set of all complete theories $\mathsf{MC}(\mathcal{B})$. The belief set $B \subseteq \mathsf{MC}(\mathcal{B})$ consists of all the maximally consistent theories which are an extension of the theory A, that is,

$$B = \{\Sigma \in \mathsf{MC}(\mathcal{B}) \mid A \subseteq \Sigma\}.$$

The interpretation $I : \mathsf{At} \to \mathcal{P}(\mathsf{MC}(\mathcal{B}))$ is defined such that an atomic sentence $p \in \mathsf{At}$ is true at all those maximally consistent theories that contain the sentence p, that is,

$$I(p) = \{\Sigma \in \mathsf{MC}(\mathcal{B}) \mid p \in \Sigma\}.$$

It is a standard argument, corresponding to the Truth Lemma in modal completeness proofs, to show that in this model for all formulas $\varphi \in \mathcal{B}$ and complete theories $\Sigma \in \mathsf{MC}(\mathcal{B})$ it holds that

$$\Sigma \in I(\varphi) \quad \text{iff} \quad \varphi \in \Sigma. \tag{7.1}$$

We use this to show that $A = A^M$.

First assume we are given a $\varphi \in A$. We want to show that then also $\varphi \in A^M$, meaning that φ is true at all worlds $\Sigma \in B$. So consider any $\Sigma \in B$. By definition of B this means that $A \subseteq \Sigma$ and hence with the assumption that $\varphi \in A$ it follows that $\varphi \in \Sigma$. By (7.1) this means that $\Sigma \in I(\varphi)$ and so we are done.

We prove that $A \supseteq A^M$ by showing for any propositional formula φ that $\varphi \notin A$ implies $\varphi \notin A^M$. From $\varphi \notin A$ it follows by Lindenbaum's Lemma that there is a complete theory $\Sigma \in \mathsf{MC}(\mathcal{B})$ with $A \subseteq \Sigma$ and $\varphi \notin \Sigma$. From the former it follows that $\Sigma \in B$ and from the latter we get by (7.1) that $\Sigma \not\models \varphi$. But now we have a world in the belief set B at which φ is false. By the definition of A^M this means that $\varphi \notin A^M$. □

Because the setting of Section 2.2 does not depend much on the representation of meanings we can treat the atomic and the propositional cases together in one single proof.

7.1.3. Proposition. *Assume that either $\mathcal{V} = \mathsf{At}$ or that $\mathcal{V} = \mathcal{B}$. A behavior $A \subseteq \mathcal{V}$ is interpretable with an interpretation $I : \mathsf{At} \to \mathcal{P}W$ iff it is closed under implications relative to I, that is, it satisfies for all sentences $\psi \in \mathcal{V}$ and set of sentences $F \subseteq \mathcal{V}$ that:*

$$\textit{If } \bigcap_{\varphi \in F} I(\varphi) \subseteq I(\psi) \textit{ and } F \subseteq A \textit{ then } \psi \in A. \tag{CI}$$

Proof. For the left-to-right direction assume that there is some belief set $B \subseteq W$ such that $A = A^M$ for the simple possible world model $M = (W, B, I)$. Take any $\psi \in \mathcal{V}$ such that $\bigcap_{\varphi \in A^M} I(\varphi) \subseteq I(\psi)$. By the definition of A^M we have that $B \subseteq I(\varphi)$ for every $\varphi \in A^M$. Hence also $B \subseteq \bigcap_{\varphi \in A^M} I(\varphi)$. By the assumption that $\bigcap_{\varphi \in A^M} I(\varphi) \subseteq I(\psi)$ it follows that $B \subseteq I(\psi)$ and so $\psi \in A^M$.

For the other direction assume we are given a behavior A that satisfies (CI). We need to find a set $B \subseteq W$ such that $A^M = A$ for the model $M = (W, B, I)$. This set is defined as follows:

$$B = \bigcap_{\varphi \in A} I(\varphi).$$

We need to show that with this definition $A^M = A$. First take a $\psi \in A^M$. By the definition of A^M this means that $B \subseteq I(\psi)$. By unfolding definition of B on sees that this is precisely the antecedent of (CI). Hence it follows that $\psi \in A$. We have that $\psi \in A^M$ for any $\psi \in A$ because $B \subseteq I(\psi)$ holds by the definition of B for any such $\psi \in A$. □

7.2 Conditions for interpretability

In this section I collect different conditions for interpretability that are defined throughout the thesis and are used in the following sections. They concern interpretability with either a belief function $b : S \to \mathcal{P}W$ or an evidence function $e : S \to \mathcal{P}W$. To abstract away from the differences between belief functions and evidence functions I formulate them as conditions for interpretability with some function $c : S \to \mathcal{P}W$, where c can then be taken to be either a belief function or an evidence function.

A behavior $a : S \to \mathcal{P}\mathcal{V}$ satisfies the *simple covering condition* relative to some function $c : S \to \mathcal{P}W$ if for every situation $s \in S$ and every set of situations $T \subseteq S$ it holds that

$$c(s) \subseteq \bigcup_{t \in T} c(t) \text{ implies } \bigcap_{t \in T} a(t) \subseteq a(s). \tag{SC}$$

A behavior $a : S \to \mathcal{P}\mathcal{V}$ satisfies the *exact covering condition* relative to some function $c : S \to \mathcal{P}W$ if for all $s \in S$ and $T \subseteq S$ it holds that

$$c(s) = \bigcup_{t \in T} c(t) \text{ implies } \bigcap_{t \in T} a(t) \subseteq a(s). \tag{EC}$$

A behavior $a : S \to \mathcal{P}\mathcal{V}$ satisfies the *monotonicity condition* relative to some function $c : S \to \mathcal{P}W$ if for all $s, t \in S$ it holds that

$$c(s) \subseteq c(t) \text{ implies } a(t) \subseteq a(s). \tag{Mon}$$

In the definition of the remaining conditions it is presupposed that the sentences in $\mathcal{V}$ can be combined by propositional connectives and that a notion of logical consequence is defined for them. This then entails that we can write $\mathsf{cl}\,(\Sigma)$ for the set of all logical consequences of Σ and we can call a set $\Sigma \subseteq \mathcal{V}$ of sentences consistent if $\bot \notin \mathsf{cl}\,(\Sigma)$.

In the case $\mathcal{V} = \mathcal{B}$, where we are working with propositional formulas, I generally assume that the notions of logical consequence and consistency are defined relative to classical propositional logic, as explained in Section 1.2.

In the case $\mathcal{V} = \mathcal{M}$, where we are working with modal formulas, one needs to explicitly state what notion of consequence and consistency one has in mind when applying one of the following conditions of interpretability.

A behavior $a : S \to \mathcal{P}\mathcal{V}$ satisfies the *conjunctive covering condition* relative to some function $c : S \to \mathcal{P}W$ if for all $s \in S$ and $s_{j,k} \in S$ for every $j \in J$, of some index set J, and $k \in \{1, \ldots, n_j\}$, for some number n_j, it holds that

$$c(s) \subseteq \bigcup_{j \in J} (c(s_{j,1}) \cap \cdots \cap c(s_{j,n_j})) \text{ implies } \bigcap_{j \in J} \mathsf{cl}\left(a(s_{j,1}) \cup \cdots \cup a(s_{j,n_j})\right) \subseteq a(s). \tag{CC}$$

A behavior $a : S \to \mathcal{PV}$ satisfies the *conjunctive consistency condition* relative to some function $c : S \to \mathcal{P}W$ if for all $s_1, \ldots, s_n \in S$ it holds that

$$c(s_1) \cap \cdots \cap c(s_n) \neq \emptyset \text{ implies that } \mathsf{cl}\left(a(s_1) \cup \cdots \cup a(s_n)\right) \text{ is consistent.} \tag{CCons}$$

A behavior $a : S \to \mathcal{PV}$ satisfies the *conjunction condition* relative to some function $c : S \to \mathcal{P}W$ if for all $s, t_1, \ldots, t_n \in S$ it holds that

$$c(s) = c(t_1) \cap \cdots \cap c(t_n) \text{ implies } \mathsf{cl}\left(a(t_1) \cup \cdots \cup a(t_n)\right) \subseteq a(s). \tag{Conj}$$

The next condition uses the notion of an n-theory from Definition 7.3.6 below. A behavior $a : S \to \mathcal{PV}$ satisfies the *cardinality condition* relative to some function $c : S \to \mathcal{P}W$ if for all $s \in S$ it holds that if $c(s)$ is finite and contains n elements then $a(s)$ is an n-theory.

To state the plausibility covering condition from Section 3.3 we need two auxiliary definitions. So fix a linguistic behavior $a : S \to \mathcal{PV}$ and a function $c : S \to \mathcal{P}W$. Consider some formula $\varphi \in \mathcal{V}$. A world $w \in W$ *is potentially* φ *in a set of worlds* $X \subseteq W$ if there are $s_1, \ldots, s_n \in S$ such that $w \in c(s_i)$ and $c(s_i) \subseteq X$ for all $i \in \{1, \ldots, n\}$ and $\varphi \in \mathsf{cl}\left(a(s_1) \cup \cdots \cup a(s_n)\right)$. A world $w \in W$ *is implausible in a set of worlds* $X \subseteq W$ if w is potentially $\bot$ in X.

A behavior $a : S \to \mathcal{PV}$ satisfies the *plausibility covering condition* relative to a function $c : S \to \mathcal{P}W$ if for all $\varphi \in \mathcal{B}$ and $s \in S$ it holds that $\varphi \in a(s)$ whenever there is an $X \subseteq W$ such that $c(s) \subseteq X$, all $w \in X \setminus c(s)$ are implausible in X and all $w \in c(s)$ are potentially φ in X.

The last condition assumes that the language of the subject contains a unary modal operator $\Box$. For instance we might have $\mathcal{V} = \mathcal{B}$ as in Chapter 6. A behavior $a : S \to \mathcal{PM}$ satisfies *that necessity is a priori* if for all situations $s, t \in S$ and sentences $\varphi \in \mathcal{M}$ it holds that:

$$\text{If } a(s) \text{ is consistent and } \Box\varphi \in a(s) \text{ then } \Box\varphi \in a(t).$$

We conclude this section by proving some entailments between the above conditions.

7.2.1. Proposition. *The simple covering condition* (SC) *entails the exact covering condition* (EC) *and the monotonicity condition* (Mon).

If $c[S] = \{c(s) \subseteq W \mid s \in S\}$ *is closed under arbitrary non-empty unions then the exact covering condition* (EC) *and the conjunction condition* (Conj) *together entail the conjunctive covering condition* (CC).

The conjunction condition (Conj) *entails the monotonicity condition* (Mon)

The conjunctive covering condition (CC) *entails the exact covering condition* (EC) *and the conjunction condition* (Conj)*.*

If $c[S] = \{c(s) \subseteq W \mid s \in S\}$ *is closed under arbitrary non-empty unions and finite non-empty intersections then the exact covering condition* (EC) *and the conjunction condition* (Conj) *together entail the conjunctive covering condition* (CC)*.*

Proof. It is obvious that (SC) entails (EC)

To see that (CC) entails (Mon) consider the case of (CC) where T is a singleton set.

To show that (SC) follows from the conjunction of (EC) and (Mon) assume that latter two and suppose that $c(s) \subseteq \bigcup_{t\in T} c(t)$. We need to show that $\bigcap_{t\in T} \mathsf{cl}\,(a(t)) \subseteq a(s)$. We can assume that T is non-empty because otherwise $c(s) = \emptyset$ and then $\mathsf{At} \subseteq b(s)$ follows by the instance of (SC) where I is the empty set. Because $c[S]$ is closed under arbitrary non-empty unions there must also be a $d \in S$ such that $c(d) = \bigcup_{t\in T} c(t)$. With (EC) we get that $\bigcap_{t\in T} a(t) \subseteq a(d)$. Because of the assumption that $c(s) \subseteq \bigcup_{t\in T} c(t)$ it follows that $c(s) \subseteq c(d)$ and hence $a(d) \subseteq a(s)$ by (Mon). Together with $\bigcap_{t\in T} a(t) \subseteq a(d)$ this yields $\bigcap_{t\in T} \mathsf{cl}\,(a(t)) \subseteq a(s)$.

To see that (Conj) entails (Mon) assume that $c(s) \subseteq c(t)$. This entails that $c(s) = c(t) \cap C(s)$. So we can apply (Conj) to obtain that $\mathsf{cl}\,(a(t) \cup a(s)) \subseteq a(s)$. It follows that $a(t) \subseteq a(s)$ because clearly $a(t) \subseteq \mathsf{cl}\,(a(t)) \subseteq \mathsf{cl}\,(a(t) \cup a(s))$.

To see that (CC) entails (EC) instantiate (CC) such that $n_j = 1$ for every $j \in J$.

To see that (CC) entails (Conj) instantiate (CC) such that J is a singleton set.

To show that (CC) follows from the conjunction of (EC) and (Conj) assume that latter two and suppose that $c(s) \subseteq \bigcup_{j\in J}(c(s_{j,1}) \cap \cdots \cap c(s_{j,n_j}))$. We need to show that $\bigcap_{j\in J} \mathsf{cl}\,\big(a(s_{j,1}) \cup \cdots \cup a(s_{j,n_j})\big) \subseteq a(s)$. We can assume that J is non-empty because otherwise $c(s) = \emptyset$ and then $\mathcal{V} \subseteq b(s)$ follows by the instance of (EC) where J is the empty set.

Because $c[S]$ is closed under finite intersections there is an $e_j \in S$ such that $c(e_j) = c(s_{j,1}) \cap \cdots \cap c(s_{j,n_j})$ for every $j \in J$. By (Conj) it follows for every $j \in J$ that $\mathsf{cl}\,\big(a(s_{j,1}) \cup \cdots \cup a(s_{j,n_j})\big) \subseteq a(e_j)$. Because $c[S]$ is closed under arbitrary non-empty unions there must also be a $d \in S$ such that $c(d) = \bigcup_{j\in J} c(e_j)$ and from (EC) we know that $\bigcap_{j\in J} a(e_j) \subseteq a(d)$. So it also follows that $c(d) = \bigcup_{j\in J}(c(s_{j,1}) \cap \cdots \cap c(s_{j,n_j}))$ and that $\bigcap_{j\in J} \mathsf{cl}\,\big(a(s_{j,1}) \cup \cdots \cup a(s_{j,n_j})\big) \subseteq a(d)$. Our assumption that $c(s) \subseteq \bigcup_{j\in J}(c(s_{j,1}) \cap \cdots \cap c(s_{j,n_j}))$ can now be stated as $c(s) \subseteq c(d)$. From this it follows by (Mon), which as show above is a consequence of (Conj), that $a(d) \subseteq a(s)$. The conclusion follows because we know from above that $\bigcap_{j\in J} \mathsf{cl}\,\big(a(s_{j,1}) \cup \cdots \cup a(s_{j,n_j})\big) \subseteq a(d)$. □

7.3 Multiple situations

This section contains the representation theorems for the notions of tight and splitting interpretability as discussed in Section 2.6 and for tight and splitting prior belief interpretability as discussed in Section 3.2.

Because I consider prior belief interpretability with an evidence function $e : S \to \mathcal{P}W$ in parallel with interpretability with a belief function $b : S \to \mathcal{P}W$ it is convenient to speak of interpretability with a function $c : S \to \mathcal{P}W$, where we might think of c as either being an evidence function e or a belief function b.

The notions of tight and splitting interpretability from Section 2.6 and of tight and splitting prior belief interpretability from Section 3.2 are all instances of a more general definition of interpretability.

7.3.1. Definition. Let $f : W' \to W$ be any function. A linguistic behavior $a : S \to \mathcal{P}\mathcal{V}$ is *f-interpretable with* $c : S \to \mathcal{P}W$ if $a = a^M$ for some multi-situation model $M = (W', b', I)$ where $b' : S \to \mathcal{P}W'$ is defined such that $b'(s) = f^{-1}[c(s)]$ for all $s \in S$.

The behavior $a : S \to \mathcal{P}\mathcal{V}$ is *tightly interpretable with* $c : S \to \mathcal{P}W$ if it is f-interpretable with c for some bijective function $f : W' \to W$.

The behavior $a : S \to \mathcal{P}\mathcal{V}$ is *splitting interpretable with* $c : S \to \mathcal{P}W$ if it is f-interpretable with c for some surjective function $f : W' \to W$.

The behavior $a : S \to \mathcal{P}\mathcal{V}$ is *tightly prior belief interpretable with* $c : S \to \mathcal{P}W$ if it is f-interpretable with c for some injective function $f : W' \to W$.

The behavior $a : S \to \mathcal{P}\mathcal{V}$ is *splitting prior belief interpretable with* $c : S \to \mathcal{P}W$ if it is f-interpretable with c for some function $f : W' \to W$.

One can easily check that the notions of interpretability defined here are equivalent to the ones from Definitions 2.6.1, 2.6.2, 3.2.2 and 3.2.4.

I first characterize all these notions in the case where $\mathcal{V} = \mathsf{At}$.

7.3.2. Theorem. *Tight and splitting interpretability and tight and splitting prior belief interpretability of a behavior $a : S \to \mathcal{P}\mathsf{At}$ with a function $c : S \to \mathcal{P}W$ all coincide and they are equivalent to the simple covering condition* (SC)*.*

Proof. It is clear from Definition 7.3.1 that tight interpretability implies the all the other notions of interpretability and that splitting prior belief interpretability is implied by all the other notions of interpretability. So it sufficient to show that splitting prior belief interpretability implies (SC) and that (SC) implies tight interpretability.

We first show that splitting prior belief interpretability implies (SC). So assume that are given a function $c : S \to \mathcal{P}W$, a model (W', b', I) and a function $f : W' \to W$. We need to show that $a^M : S \to \mathcal{P}\mathsf{At}$ satisfies (SC). So assume we have an $s \in S$ and a $T \subseteq S$ such that $c(s) \subseteq \bigcup\{c(t) \mid t \in T\}$. Because $f^{-1}[\cdot] : \mathcal{P}W \to \mathcal{P}W'$ preserves inclusions and arbitrary unions and by definition $b'(s) = f^{-1}[c(s)]$ for any $s \in S$ it follows that $b'(s) \subseteq \bigcup\{b'(t) \mid t \in T\}$.

Now pick any atomic sentence $p \in \bigcap\{a^M(t) \mid t \in T\}$. We need to show that $b'(s) \subseteq I(p)$ because this yields that $p \in a^M(s)$. Since $b'(s) \subseteq \bigcup\{b'(t) \mid t \in T\}$ it suffices to show that $b'(t) \subseteq I(p)$ for every $t \in T$. But this holds because for any $t \in T$ we have that $p \in a^M(t)$.

We now prove that whenever a behavior $a : S \to \mathcal{P}\mathsf{At}$ and a function $c : S \to \mathcal{P}W$ satisfy (SC) then a is tightly interpretable with c. To show that a is tightly interpretable with c it suffices to find an interpretation $I : \mathsf{At} \to \mathcal{P}W$ such that $a = a^M$ for the model $M = (W, c, I)$.

Define the interpretation function I such that

$$I(p) = \bigcup\{c(s) \subseteq W \mid \text{for some } s \in S \text{ with } p \in a(s)\}.$$

To show that $a = a^M$ we need to prove that for all $s \in S$ and $p \in \mathsf{At}$

$$p \in a(s) \quad \text{iff} \quad c(s) \subseteq I(p).$$

The left-to-right direction follows immediately from the definition of I.

For the right-to-left direction assume that $c(s) \subseteq I(p)$. We need to show that $p \in a(s)$. The idea is to find a suitable covering of $c(s)$ such that we can employ (SC). To do so consider any $w \in c(s)$. Because $c(s) \subseteq I(p)$ it follows from the definition of I that $w \in c(s_w)$ for some $s_w \in S$ such that $p \in a(s_w)$. Define T to be the set of all such s_w, that is, $T = \{s_w \mid w \in c(s)\}$. Because $w \in c(s_w)$ for all $w \in c(s)$ we have that $c(s) \subseteq \bigcup_{t \in T} c(t)$. It also holds that $p \in \bigcap_{t \in T} a(t)$ because we had $p \in a(s_w)$ for any $w \in c(s)$. Hence we can apply (SC) to conclude that $p \in a(s)$. □

We obtain a corollary for the single situation case from Section 2.3.

7.3.3. Corollary. *Every behavior $A \subseteq \mathsf{At}$ is tightly interpretable with a non-empty $B \subseteq W$. A behavior $A \subseteq \mathsf{At}$ is tightly interpretable with the empty set $\emptyset \subseteq W$ iff $A = \mathsf{At}$.*

Proof. Apply Theorem 7.3.2 in the case where $S = \{s\}$ is a singleton set and $b(s) = B$.

To see hat the first claim follows observe that if $b(s)$ is non-empty then (SC) is trivially satisfied because the only covering of $b(s)$ is given by the set $T = \{s\}$.

To see that the second claim follows note that if $b(s)$ is empty then we can cover it with $T = \emptyset$ and hence it follows by (SC) that $\mathsf{At} \subseteq \bigcap \emptyset \subseteq a(s)$. □

In the following we consider the case where $\mathcal{V} = \mathcal{B}$. In this setting we have to treat tight and splitting interpretability separately. I first consider the simpler setting of splitting interpretability.

7.3.4. Theorem. *A behavior $a : S \to \mathcal{P}\mathcal{B}$ is splitting prior belief interpretable with a $c : S \to \mathcal{P}W$ iff it satisfies the conjunctive covering condition* (CC).

The behavior a is splitting interpretable with c iff it satisfies the conjunctive covering condition (CC) *and the conjunctive consistency condition* (CCons).

Proof. We begin with the left-to-right direction of both claims.

First assume that $a : S \to \mathcal{PB}$ is splitting prior belief interpretable with $c : S \to \mathcal{P}W$. We need to check that then (CC) holds. That a is splitting prior belief interpretable means that it is equal to a^M for some multi-situation model $M = (W', b', I)$ where $b'(s) = f^{-1}[c(s)]$ for some function $f : W' \to W$.

Now assume that we are given $s \in S$ and $s_{j,k} \in S$ such that the antecedent of (CC) is satisfied. Moreover assume we have a $\varphi \in \mathcal{B}$ such that for all $j \in J$ it holds that $\varphi \in \mathsf{cl}\left(a^M_K(s_{j,1}) \cup \cdots \cup a^M_K(s_{j,n_j})\right)$. We want to show that then $\varphi \in a^M_K(s)$.

So we need that $w \in I(\varphi)$ for any $w \in b'(s)$. By the fact that f^{-1} preserves intersections, unions and inclusions it follows from the antecedent of (CC) that $b'(s) \subseteq \bigcup_{j \in J}(b'(s_{j,1}) \cap \cdots \cap b'(s_{j,n_j}))$. So there is some $j \in J$ such that $w \in b'(s_{j,k})$ for all $k \in \{1, \ldots, n_j\}$. Because $\varphi \in \mathsf{cl}\left(a^M(s_{j,1}) \cup \cdots \cup a^M(s_{j,n_j})\right)$ there exists a set $\Sigma \subseteq a^M(s_{j,1}) \cup \cdots \cup a^M(s_{j,n_j})$ such that $\Sigma \models \varphi$. For every $\psi \in \Sigma$ we have that $\psi \in a^M(j,k)$ for some $k \in \{1, \ldots, n_j\}$ and hence by the definition of a^M we get that $b(s_{j,k}) \subseteq I(\psi)$. Because $w \in b(s_{j,k})$ for all such k it follows that $w \in I(\psi)$ for every $\psi \in \Sigma$. By the definition of logical consequence this entails $w \in I(\varphi)$.

Now consider the left-to-right direction of the second claim. Since splitting interpretability with c implies splitting prior belief interpretability with c we have by the argument above that (CC) is satisfied whenever a is splitting interpretable with c.

Additionally, we need to check that $\mathsf{cl}\left(a^M(s_1) \cup \cdots \cup a^M(s_n)\right)$ is consistent whenever $c(s_1) \cap \cdots \cap c(s_n) \neq \emptyset$ for some $s_1, \ldots, s_n \in S$. So assume we have $a = a^M$ for a model $M = (W', b', I)$ as above with the difference that now we can assume that the function $f : W' \to W$ is surjective.

So suppose that $c(s_1) \cap \cdots \cap c(s_n) \neq \emptyset$. This means that there is some $w \in W$ such that $w \in c(s_k)$ for all $k \in \{1, \ldots, n\}$. By the subjectivity of f there is a $w' \in W'$ such that $f(w') = w$. Because $b'(s_k) = f^{-1}[c(s_k)]$ we then also have that $w' \in b'(s_k)$ for every $k \in \{1, \ldots, n\}$.

Now assume for a contradiction that $\bot \in \mathsf{cl}\left(a^M(s_1) \cup \cdots \cup a^M(s_n)\right)$. It then follows that there is a $\Sigma \subseteq a^M(s_1) \cup \cdots \cup a^M(s_n)$ such that $\Sigma \models \bot$. Because $w' \in b'(s_k)$ for all $k \in \{1, \ldots, n\}$ we then have by the definition of a^M that $w' \in I(\psi)$ for all $\psi \in \Sigma$. Hence it follow that $w' \in I(\bot)$ which is impossible.

For the right-to-left direction of the first claim assume that a satisfies (CC). We construct a model $M = (W', b', I)$ with a function $f : W' \to W$ such that $a = a^M$ and $b'(s) = f^{-1}[c(s)]$ for all $s \in S$.

The interpreting model uses the domain

$$W' = \{(w, \Sigma) \in W \times \mathsf{MC}(\mathcal{B}) \mid a(s) \subseteq \Sigma \text{ for all } s \in S \text{ with } w \in c(s)\}.$$

The interpretation I is defined such that $I(p) = \{(w, \Sigma) \in W' \mid p \in \Sigma\}$. It is easy to check that this definition implies that for all propositional formulas $\varphi \in \mathcal{B}$

$$(w, \Sigma) \in I(\varphi) \quad \text{iff} \quad \varphi \in \Sigma. \tag{7.2}$$

The function $f : W' \to W$ is the projection on the first component such that $f(w, \Sigma) = w$ for all $(w, \Sigma) \in W'$.

We need to show that for all $s \in S$ and $\varphi \in \mathcal{B}$

$$\varphi \in a(s) \quad \text{iff} \quad \varphi \in a^M(s).$$

First assume that $\varphi \in a(s)$. To show that $\varphi \in a^M(s)$ we need that $(w, \Sigma) \in I(\varphi)$ for every $(w, \Sigma) \in b(s) = f^{-1}[c(s)]$. So take any such (w, Σ). By the definition of f it follows that $w \in c(s)$. Since $(w, \Sigma) \in W'$ it follows that $a(s) \subseteq \Sigma$. Hence $\varphi \in \Sigma$ from which it follows by (7.2) that $(w, \Sigma) \models \varphi$.

For the other direction take any $\varphi \in a^M(s)$. We need to establish that $\varphi \in a(s)$. We show that for every $w \in c(s)$ we can find $s_1, \ldots, s_n \in S$ such that $w \in c(s_1) \cap \cdots \cap c(s_n)$ and $\varphi \in \mathsf{cl}\,(a(s_1) \cup \cdots \cup a(s_n))$. This then provides the required covering for an application of (CC) which then yields that $\varphi \in a(s)$.

So fix any $w \in c(s)$. We show that

$$\varphi \in \mathsf{cl}\left(\bigcup\{a(s') \mid s' \in S, w \in c(s')\}\right).$$

The claim, which involves only finitely many such s', follows by compactness.

Assume for a contradiction that $\varphi \notin \mathsf{cl}\,(\bigcup\{a(s') \mid w \in c(s')\})$. This means that $\mathsf{cl}\,(\bigcup\{a(s') \mid w \in c(s')\})$ is a consistent theory not containing φ. By Lindenbaum's Lemma it can be extended to a complete theory $\Sigma \in \mathsf{MC}(\mathcal{B})$ such that $\varphi \notin \Sigma$ and $a(s') \subseteq \Sigma$ for all s' with $w \in c(s')$. With the latter conjunct we get that $(w, \Sigma) \in W'$. Because $f(w, \Sigma) = w \in c(s)$ it holds that $(w, \Sigma) \in f^{-1}[b(s)] = b'(s)$. We also have that $b'(s) \subseteq I(\varphi)$ because of the assumption that $\varphi \in a^M(s)$. Hence $(w, \Sigma) \in I(\varphi)$ which by (7.2) entails that $\varphi \in \Sigma$. But this contradicts $\varphi \notin \Sigma$ which is guaranteed by the construction of Σ.

For the right-to-left direction of the second claim we use the same construction as for the right-to-left direction of the first claim. However now we need to show additionally that $f : W' \to W$ is surjective if (CCons) is satisfied. So assume this and pick any $w \in W$. We need to find an element in W' which maps to w under f. Because f is the projection on the first component this means that we have to find a $\Sigma \in \mathsf{MC}(\mathcal{B})$ such that $(w, \Sigma) \in W'$. To do so we show that $\mathsf{cl}\,(\bigcup\{a(s') \mid w \in c(s')\})$ is a consistent theory. This proves the claim because if $\mathsf{cl}\,(\bigcup\{a(s') \mid w \in c(s')\})$ is consistent we can use Lindenbaum's Lemma, with $\bot$ as φ, to extend it to a maximal consistent $\Sigma \in \mathsf{MC}(\mathcal{B})$ with the property that $a(s') \subseteq \Sigma$ for every $s' \in S$ such that $w \in c(s')$. This property is precisely what is required for $(w, \Sigma) \in W'$.

It remains to show that $\mathsf{cl}\,(\bigcup\{a(s') \mid w \in c(s')\})$ is consistent. So suppose for a contradiction that $\bot \in \mathsf{cl}\,(\bigcup\{a(s') \mid w \in c(s')\})$. By compactness it follows that there are already finitely many $s_1, \dots, s_l \in S$ with $w \in c(s_1) \cap \dots \cap c(s_l)$ such that $\bot \in \mathsf{cl}\,(a(s_1) \cup \dots \cup a(s_l))$. Hence we have a counterexample to our assumption that (CCons) holds. □

We obtain an analogous result for the single situation case as a corollary.

7.3.5. Corollary. *A behavior $A \subseteq \mathcal{B}$ is splitting interpretable with some non-empty $B \subseteq W$ iff A is a theory. A behavior $A \subseteq \mathcal{B}$ is splitting interpretable with the empty set $\emptyset \subseteq W$ iff $A = \mathcal{B}$.*

Proof. Apply Theorem 7.3.4 in the case where $S = \{s\}$ is a singleton set and $b(s) = B$.

To see hat the first claim follows observe that if $b(s)$ is non-empty then (CC) is trivially satisfied because the only covering of $b(s)$ is given by the set $T = \{s\}$.

To see that the second claim follows note that if $b(s)$ is empty then we can cover it with $T = \emptyset$ and hence it follows by (CC) that $\mathcal{B} \subseteq \bigcap \emptyset \subseteq a(s)$. □

We now consider tight interpretability. The difficulty with tight interpretability is that we need to ensure that the worlds in W suffice to give counterexamples to all the sentences that the subject does not accept. For this we need that the behavior of the subject satisfies the cardinality condition which constrains the set of sentences that the subject accepts in some situation according to the number of elements in her belief set of this situation.

If the belief set of the subject is a finite set with n elements then we need that the set of sentences that she accepts is an n-theory according to the following definition:

7.3.6. Definition. A theory $\Sigma \subseteq \mathcal{B}$ is a *n-theory* for some natural number $n \in \omega$ if for every finite set $F \subseteq \mathcal{B}$ with $\bigvee F \in \Sigma$ there is a subset $G \subseteq F$ such that $|G| \leq n$ and $\bigvee G \in \Sigma$.

One can easily see that the only 0-theory is the inconsistent theory because $\bigvee \emptyset = \bot$. By instantiating the definition one can see that 1-theories are precisely the theories $\Sigma \subseteq \mathcal{B}$ such that if $\varphi \vee \psi \in \Sigma$ then $\varphi \in \Sigma$ or $\psi \in \Sigma$. Such theories are called prime theories and it is well-known that in the case of classical logic they are precisely the complete theories.

The following Lemma shows an n-theory can be seen as the set of sentences that are true on a set of n many possible worlds. The converse of the Lemma is also true but we do not bother proving it here.

7.3.7. Lemma. *If a theory $\Sigma \subseteq \mathcal{B}$ is an n-theory then the set*

$$E^\Sigma = \{\Gamma \in \mathsf{MC}(\mathcal{B}) \mid \Sigma \subseteq \Gamma\}$$

contains at most n elements.

Proof. We prove the contrapositive. So assume that Σ is such that E^{Σ} contains $n+1$-many distinct theories $\Gamma_1, \dots, \Gamma_n, \Gamma_{n+1} \in E^{\Sigma}$. Because these theories are distinct we can construct for every $k \in \{1, \dots, n+1\}$ a formula $\varphi_k \in \Gamma_k$ such that $\varphi_k \notin \Gamma_l$ for all $l \neq k$. Let $F = \{\varphi_k \mid k \in \{1, \dots, n+1\}\}$. Now clearly $\bigvee F \in \Sigma$. Also $|F| = n+1$ because $\varphi_k \neq \varphi_l$ whenever $k \neq l$ since $\varphi_k \in \Gamma_k$ but $\varphi_l \notin \Gamma_k$. To prove that Σ is not an n-theory we need to show that for any subset $D \subseteq F$ such that $|D| \leq n$ it is not the case that $\bigvee D \in \Sigma$. So some any such D. Because D contains at most n elements but F contains $n+1$ elements there is at least one $k \in \{i, \dots, n+1\}$ such that $\varphi_k \notin D$. But from this it follows that $\bigvee D \notin \Gamma_i$ and hence $\bigvee D \notin \Sigma$. □

I could not find a good notion of a κ-theory that works well for infinite cardinalities κ. I avoid this problem by assuming that the set of sentences in the vocabulary of the subject is at most countable. The next lemma shows that in that case it always suffices to have countably many worlds to give counterexamples to all the sentences that the subject does not accept.

7.3.8. Lemma. *Assume that $\mathcal{B}$ is generated from a finite or at most countably infinite set* At *of atomic sentences. For every propositional theory $\Sigma \subseteq \mathcal{B}$ there exists a countable set $C \subseteq \mathsf{MC}(\mathcal{B})$ such that $\Sigma = \bigcap C$.*

Proof. Let $Z \subseteq \mathcal{B}$ set of all sentences that are not in Σ, that is, $Z = \{\varphi \in \mathcal{B} \mid \varphi \notin \Sigma\}$. The set $\mathcal{B}$ is countable because we assume At to be at most countable. Because $Z \subseteq \mathcal{B}$ it follows that also the set Z is at most countable.

Now consider a $\varphi \in Z$. Because $\varphi \notin \Sigma$ there exists by Lindenbaum's Lemma a maximal consistent theory Γ_φ with $\Sigma \subseteq \Gamma_\varphi$ and $\varphi \notin \Gamma_\varphi$. Let $C = \{\Gamma_\varphi \in \mathsf{MC}(\mathcal{B}) \mid \varphi \in Z\}$. This C is at most countable since Z is at most countable. The inclusion $\Sigma \subseteq \bigcap C$ holds because $\Sigma \subseteq \Gamma_\varphi$ for all $\varphi \in Z$. To see that also the other inclusion $\Sigma \supseteq \bigcap C$ holds consider take some $\varphi \notin \Sigma$. Then also $\varphi \notin \bigcap C$ because $\varphi \in Z$ and by construction $\varphi \notin \Gamma_\varphi \supseteq \bigcap C$. □

The previous two lemmas yield the following more general lemma:

7.3.9. Lemma. *Let $\mathcal{B}$ be based on an at most countably infinite set* At *of atomic sentences. Take some set X and a consistent theory Σ such that Σ is an n-theory if X is a finite set with n elements. Then there exists a function $g_X : X \to \mathsf{MC}(\mathcal{B})$ such that*

$$\Sigma = \bigcap g_X[X] = \bigcap \{g_X(x) \in \mathsf{MC}(\mathcal{B}) \mid x \in X\}.$$

Proof. First consider the case where X is finite and contains n elements. First observe that $E^{\Sigma} = \{\Gamma \in \mathsf{MC}(\mathcal{B}) \mid \Sigma \subseteq \Gamma\}$ is non-empty. Because Σ is consistent we have that $\bot \notin \Sigma$. By Lindenbaum's Lemma it follows that there is a maximal consistent set of formulas extending Σ. Moreover, we can apply Lemma 7.3.7 because Σ is an n-theory. It follows that E^{Σ} is a non-empty set with less elements than X. Hence there exists a surjective function $g'_X : X \to E^\Sigma$ which can also be seen as a function $g_X : X \to \mathsf{MC}(\Sigma)$ such that $g_X[X] = E^\Sigma$.

It remains to show that $\Sigma = \bigcap E^\Sigma$. The $\subseteq$-inclusion holds because by definition Σ is contained in each of the elements of E^Σ. For the $\supseteq$-inclusion take any $\varphi \notin \Sigma$. Then $\varphi \notin \bigcap E^\Sigma$ because by Lindenbaum's Lemma there exists a maximal consistent theory Γ with $\Sigma \subseteq \Gamma$ such that $\varphi \notin \Gamma$.

In the other case where X is infinite we use Lemma 7.3.8 to get an at most countably infinite set $C \subseteq \mathsf{MC}(\mathcal{B})$ such that $\Sigma = \bigcap C$. This C must be non-empty because otherwise $\Sigma = \bigcap C$ would be inconsistent. Because countable infinity is the smallest infinite cardinality and C is non-empty there exists a surjective function $g'_X : X \to C$. This gives us a function $g_X : X \to \mathsf{MC}(\mathcal{B})$ such that $g_X[X] = C$, which entails that $\Sigma = \bigcap C = \bigcap g_X[X]$. □

We can now proof the representation results for tight interpretability. It however presupposes implausible closure conditions for $c[S]$. I do not know of any concise characterization of tight interpretability without this closure condition.

7.3.10. THEOREM. *If a behavior $a : S \to \mathcal{PB}$ is tightly prior belief interpretable with a function $c : S \to W$ then it satisfies the conjunctive covering condition* (CC) *and the cardinality condition relative to c.*

If a behavior $a : S \to \mathcal{PB}$ is tightly interpretable with a function $c : S \to W$ then it satisfies the conjunctive covering condition (CC)*, the conjunctive consistency condition* (CCons)*, and the cardinality condition relative to c.*

The converse of both claims holds under the assumptions that there are finitely or at most countably infinitely many atomic sentences in At *and that the set $c[S] = \{c(s) \subseteq W \mid s \in S\}$ is closed under arbitrary non-empty intersections and under complements.*

Proof. We first prove the left-to-right directions of both claims. Since tight prior belief interpretability and tight interpretability both entail splitting prior belief interpretability it follows by Theorem 7.3.4 that they both entail the conjunctive covering condition. Because tight interpretability entails splitting interpretability it also follows by Theorem 7.3.4 that tight interpretability entails the conjunctive consistency condition.

It remains to show that if a is tightly prior belief interpretable with c then it satisfies the cardinality condition relative to c. This also covers the case of tight interpretability because tight interpretability entails tight prior belief interpretability.

So consider the behavior $a^M : S \to \mathcal{P}W$ generated by some model $M = (W', b, I)$ where $b'(s) = f^{-1}[c(s)]$ for some injective function $f : W' \to W$. We need to show that for every $s \in S$ the set $a^M(s)$ is an n-theory whenever $c(s)$ contains a finite number of n elements.

To do this assume that $c(s)$ contains n elements. We need to show that $a^M(s)$ is an n-theory. For this pick any finite set of formulas $F \subseteq \mathcal{B}$ with $\bigvee F \in a^M(s)$.

By the definition of a^M the latter means that $b'(s) = f^{-1}[c(s)] \subseteq I(\bigvee F)$. Because $c(s)$ has cardinality n, $b'(s) = f^{-1}[c(s)]$ and f is injective it follows that $b'(s)$ has a cardinality smaller or equal n. Since $b'(s) \subseteq I(\bigvee F)$ we know that for every $w \in b'(s)$ there is a $\varphi_w \in F$ such that $w \in I(\varphi_w)$. Let $D = \{\varphi_w \mid w \in b'(s)\}$. Clearly $w \in I(\bigvee D)$ for every $w \in b'(s)$ and hence $\bigvee D \in a^M(s)$. Also D is a subset of F with a cardinality less or equal n because $b'(s)$ has a cardinality less or equal n. So we have found the required set D to witness that $a^M(s)$ is an n-theory.

We first give a proof of the converse of the first claim and then show how to extend it to a proof of the converse of the second claim. So assume that we are given a linguistic behavior $a : S \to \mathcal{PB}$ and a function $c : S \to \mathcal{P}W$ which satisfy the conjunctive covering condition (CC) and the cardinality condition. We construct a model $M = (W', b', I)$ for some $W' \subseteq W$ and with $b'(s) = W' \cap c(s)$ for all $s \in S$ such that $a = a^M$. This shows that a is tightly prior belief interpretable with c because we can take the inclusion of the set W' into the set W and the injective function f that is mentioned in Definition 7.3.1.

For every $w \in W$ we define the set $N(w) \subseteq W$ as

$$N(w) = \bigcap \{c(s) \subseteq W \mid w \in c(s)\}.$$

Let $\mathcal{N} = \{N(w) \mid w \in W\}$. Because $\{c(s) \mid s \in S\}$ is closed under intersections we can choose for every $N \in \mathcal{N}$ an $s_N \in S$ such that $N = c(s_N)$.

We can show that if $v \in N(w)$ then $N(v) = N(w)$. Assume $v \in N(w)$. Then $v \in c(s_{N(w)})$ and hence by the definition of $N(v)$ it follows that $N(v) \subseteq c(s_{N(w)}) = N(w)$. To see that also $N(w) \subseteq N(v)$ we show that $w \in N(v)$. From this $N(w) \subseteq N(v)$ follows with an argument analogous to the one for the other inclusion above. To see that $w \in N(v)$ we derive a con tradition from $w \notin N(v)$. Assume $w \notin N(v)$. Because $c[S]$ is closed under complements there is a $t \in S$ such that $c(t) = W \setminus c(s_{N(v)}) = W \setminus N(v)$. Because $w \notin N(v)$ it follows that $w \in c(t)$ and hence $N(w) \subseteq c(t)$ by the definition of $N(w)$. But because $v \in N(w)$ we then have that $v \in c(t)$ which contradicts $v \in N(v)$ because $c(t)$ is the complement of $N(v)$.

It now follows that the elements of $\mathcal{N}$ are mutually disjoint. Assume that $N(w) \cap N(v) \neq \emptyset$. Then there is a $u \in W$ such that $u \in N(w)$ and $u \in N(v)$. From the former it follows that $N(u) = N(w)$ and form the latter that $N(u) = N(v)$, which entails $N(w) = N(v)$.

We continue by defining the domain of the model as

$$W' = \bigcup \{N \in \mathcal{N} \mid a(s_N) \text{ is consistent}\}.$$

Now consider any $N \in \mathcal{N}$ such that $a(s_N)$ is consistent. Because of the cardinality condition it holds that $a(s_N)$ is a n-theory whenever $c(s_N)$ contains a finite number of n elements. Hence we can apply Lemma 7.3.9 to obtain a function $g_N : N \to$

$\mathsf{MC}(\mathcal{B})$ such that $a(s_N) = \bigcap g_N[N]$. Because the different N are disjoint we can paste all those functions g_N together and obtain one function $g : W' \to \mathsf{MC}(\mathcal{B})$. This g has the property that $a(s_N) = \bigcap g[N]$ for every $N \in \mathcal{N}$ such that $a(s_N)$ is consistent.

We define the interpretation $I : \mathsf{At} \to \mathcal{P}W$ such that

$$I(p) = \{w \in W' \mid p \in g(w)\}.$$

By an induction we can proof that this definition makes it the case that for all $w \in W'$

$$w \in I(\varphi) \quad \text{iff} \quad \varphi \in g(w). \tag{7.3}$$

We need to show that $\varphi \in a(s)$ iff $\varphi \in a^M(s)$.

So pick a $\varphi \in a(s)$. To conclude that $\varphi \in a^M(s)$ we need to check that $w \in I(\varphi)$ for every $w \in b'(s)$. So pick any $w \in b'(s) = W' \cap c(s)$. Let $N = N(w)$. By the definition of $N(w)$ we have that $N \subseteq c(s)$. Also $N = c(s_N)$ and hence $c(s_N) \subseteq c(s)$. By the monotonicity condition (Mon), which by Proposition 7.2.1 is a consequence of (CC), it follows that $a(s) \subseteq a(s_N)$. Our assumption is that $\varphi \in a(s)$ and so we obtain $\varphi \in a(s_N)$. Now consider $g(w) \in \mathsf{MC}(\mathcal{B})$. Because $a(s_N) = \bigcap g[N]$ and $g(w) \in g[N]$ we have that $a(s_N) \subseteq g(w)$. Hence $\varphi \in g(w)$ and by (7.3) it follows that $w \in I(\varphi)$.

For the other inclusion take any $\varphi \in a^M(s)$. We want to show that $\varphi \in a(s)$. We show that $\varphi \in a(s)$ follows from (CC) by constructing a suitable covering of $c(s)$. First observe that

$$c(s) \subseteq \bigcup\{c(s_{N(w)}) \mid w \in c(s)\}.$$

This holds because $w \in N(w) = c(s_{N(w)})$ for each $w \in c(s)$. This already provides the antecedent for applying (CC) with an instance where all intersections are intersections of singletons. To conclude $\varphi \in a(s)$ it remains to show that $\varphi \in a(s_{N(w)})$ for every $w \in c(s)$.

So assume for a contradiction that $\varphi \notin a(s_{N(w)})$ for some $w \in c(s)$. Let $N = N(w) = c(s_{N(w)})$. From $\varphi \notin a(s_N)$ and $a(s_N) = \bigcap g[N] = \bigcap\{g(w') \mid w' \in N\}$ it follows that there is a $w' \in N$ such that $\varphi \notin g(w')$. By (7.3) this means that $w' \notin I(\varphi)$. This is a contradiction to $\varphi \in a^M(s)$ because we can show that $w' \in c(s)$. The latter follows because $w' \in N = N(w)$ and by the definition of $N(w)$ it holds that $N(w) \subseteq c(s)$.

For the converse of the second claim we show that we can take $W' = W$ in the above proof if we additionally assume the conjunctive consistency condition (CCons). By the definition of W' this amounts to the claim that $a(s_{N(w)})$ is consistent for every $w \in W$. But since $w \in N(w) = c(s_{N(w)})$ this follows as the singleton instance from (CCons). □

We obtain the following corollary for the single situation case:

7.3.11. Corollary. *Assume that* At *is finite or countable. A behavior* $A \subseteq \mathcal{B}$ *is tightly interpretable with some* $B \subseteq W$ *iff* A *is a theory that is an* n*-theory whenever* B *is a finite set with* n *elements.*

Proof. Apply Theorem 7.3.10 in the case where $S = \{s\}$ is a singleton set and $b(s) = B$.

To see hat the first claim follows observe that if $b(s)$ is non-empty then (CC) is trivially satisfied because the only covering of $b(s)$ is given by the set $T = \{s\}$. If $b(s)$ is empty the $a(s)$ needs to be inconsistent because we want to satisfy (CC) for the covering $T = \emptyset$. But it already follows that $a(s)$ is inconsistent because by the cardinality condition $a(s)$ is a 0-theory whenever $b(s)$ is empty. □

7.4 Plausibility orders

In this section I consider the notion of plausibility interpretability as discussed in Section 3.3.

To prove representation theorems for plausibility interpretability we need to understand the properties of the minimization in a well-founded preorder that occurs in the definition of plausibility interpretability. The formal properties of this minimization procedure are rather complicated. I do not prove all the necessary results in this section but refer instead to (Marti and Pinosio 2016). In the following I describe how the problem of characterizing plausibility interpretability relates to the results from this paper.

The results from (Marti and Pinosio 2016) are formulated in terms of the conditionals that are true in some well-founded preorder. A conditional is a pair $(A, C) \in \mathcal{P}W \times \mathcal{P}W$ of propositions over some domain W. We write the conditional (A, C) as $A \mathrel{|\!\sim} C$. We define a conditional $A \mathrel{|\!\sim} C$ to be true on some well-founded preorder $\leq \, \subseteq W \times W$ over W if all the $\leq$-minimal A-worlds are C-worlds. We also write $\leq \, \models A \mathrel{|\!\sim} C$ if $A \mathrel{|\!\sim} C$ is true on $\leq$. So we have that

$$\leq \, \models A \mathrel{|\!\sim} C \quad \text{iff} \quad \mathsf{Min}_{\leq}(A) \subseteq C.$$

Let me explain how the notion of splitting interpretability relates to the conditionals that are true in some order. The definition of splitting plausibility interpretability from Definition 3.3.2 can be reformulated as follows: A linguistic behavior $a : S \to \mathcal{P}\mathcal{V}$ is plausibility interpretable with $e : S \to \mathcal{P}W$ if and only if there is some function $f : W' \to W$, a well-founded preorder $\leq$ on W' and an interpretation function $I : \mathsf{At} \to \mathcal{P}W'$ such that for all $s \in S$ and $\varphi \in \mathcal{V}$

$$\varphi \in a(s) \quad \text{iff} \quad \leq \, \models f^{-1}[e(s)] \mathrel{|\!\sim} I(\varphi). \tag{7.4}$$

Finding a characterization of splitting plausibility interpretability amounts roughly to finding a condition on linguistic behaviors that guarantee the existence of a well-founded preorder that makes the conditionals that are required by

(7.4) true. To find such a condition and prove that it is equivalent to splitting plausibility interpretability we need to understand the properties of those sets of conditionals that arise as the set of conditionals that are true on some order.

The properties of the sets of conditionals that are true on some well-founded preorder can be captured by a formal proof system that is called system P_∞. It concerns inferences of the form $\Phi/A \mathrel{\sim\!\!\!\!|} C$ where $A, C \subseteq W$ and Φ is a set of conditionals between propositions over W. The results from (Marti and Pinosio 2016) show that, roughly speaking, a set of conditionals Φ is the set of conditionals that is true in some well-founded preorder if and only if Φ is closed under provability in system P_∞, that is, $A \mathrel{\sim\!\!\!\!|} C \in \Phi$ whenever the inference $\Phi/A \mathrel{\sim\!\!\!\!|} C$ is provable in system P_∞.

The rules of the proof system P_∞ can be found in (Marti and Pinosio 2016, sec. 2), but they are not needed for proving the results of this section. Instead I am using the following characterization of provability in the system P_∞, which is an immediate consequence of Theorem 17 in (Marti and Pinosio 2016):

7.4.1. Theorem. *An inference $\Phi/A \mathrel{\sim\!\!\!\!|} C$ is provable in system P_∞ iff there is a $\Psi \subseteq \Phi$ such that the following two conditions are satisfied:*

1. $A \subseteq C \cup U(\Psi)$.

2. $B \cap D \subseteq (A \cap C) \cup U(\Psi)$ *for all* $B \mathrel{\sim\!\!\!\!|} D \in \Psi$.

Here $U(\Psi) \subseteq W$ is defined to be the set $U(\Psi) = \bigcup\{B \setminus D \mid B \mathrel{\sim\!\!\!\!|} D \in \Psi\}$.

Note that from this theorem it especially follows that if $A \mathrel{\sim\!\!\!\!|} C \in \Phi$ then $\Phi/A \mathrel{\sim\!\!\!\!|} C$ is provable in P_∞ because we can satisfy the above condition by choosing $\Psi = \{A \mathrel{\sim\!\!\!\!|} C\}$.

The connection between provability in P_∞ and truth in well-founded preorders is established by a soundness and a completeness theorem. Using the presentation of system P_∞ from (Marti and Pinosio 2016, sec. 2) it is easy to prove the following soundness theorem:

7.4.2. Theorem. *Let Φ be a set of conditionals over W and $A, C \subseteq W$ such that $\Phi/A \mathrel{\sim\!\!\!\!|} C$ is provable in P_∞. Take a well-founded preorder $\leq$ on W such that every $B \mathrel{\sim\!\!\!\!|} D \in \Phi$ is true in $\leq$. Then $A \mathrel{\sim\!\!\!\!|} C$ is also true in $\leq$.*

The version of the completeness theorem that we need is Corollary 21 from (Marti and Pinosio 2016). It is as follows:

7.4.3. Theorem. *Let Φ be a set of conditionals over a set of worlds W. Then there is a set of worlds W', a function $f : W' \to W$ and a well-founded poset $\leq$ on W' such that for all $A, C \subseteq W$ the inference $\Phi/A \mathrel{\sim\!\!\!\!|} C$ is provable in system P_∞ iff $f^{-1}[A] \mathrel{\sim\!\!\!\!|} f^{-1}[C]$ is true in $\leq$.*

Intuitively, this theorem states that whenever a set of conditionals is closed under inferences in P_∞ then it is the set of conditionals that are true in some well-founded preorder, modulo the duplication of worlds that is given by the function $f : W' \to W$. Here I do not explain why we need the $f : W' \to W$ and can not take $\leq$ to be defined directly on W. The interested reader is referred to Remark 5 in (Marti and Pinosio 2016) and the further references given there.

Let me know explain the strategy behind the proofs of the representation results for plausibility interpretability. Assume we need to show that a linguistic behavior is splitting plausibility interpretable if and only if it satisfies a certain condition C.

For the left-to-right direction we assume that a linguistic behavior is plausibility interpretable and want to show that it satisfies the condition C. Because the behavior is plausibility interpretable it follows that there is some well-founded preorder for which (7.4) holds. To show that the linguistic behavior satisfies the condition C we use (7.4) to translate the condition C into a closure property of the set of conditionals that are true in the well-founded preorder. This property, stated in terms of conditionals, then follows because we know from Theorem 7.4.2 that the set of conditionals that are true in an order is closed under provability in system P_∞.

For the right-to-left direction we are given a linguistic behavior satisfying the condition C. We have to find an interpreting well-founded preorder. To this aim we define a suitable set of conditionals that contains all the conditionals that according to (7.4) need to be true on an interpreting order. We then show that condition C guarantees that this set of conditionals is closed under provability in system P_∞. This allows to use completeness as stated in Theorem 7.4.3 to obtain a plausibility order that satisfies (7.4) and hence interprets the linguistic behavior.

We first prove the representation theorem in the case where $\mathcal{V} = \mathsf{At}$.

7.4.4. Theorem. *A linguistic behavior $a : S \to \mathcal{P}\mathsf{At}$ is splitting plausibility interpretable with an evidence function $e : S \to \mathcal{P}W$ iff it satisfies the exact covering condition* (EC)*.*

Proof. First assume that the behavior $a : S \to \mathcal{P}\mathsf{At}$ is plausibility interpretable with the evidence function $e : S \to \mathcal{P}W$. We show that then (EC) is satisfied. So let $M = (W', b', I)$ be some multi-situation model, $f : W' \to W$ some function and $\leq$ a well-founded preorder on W' such that $a = a^M$ and $b'(s) = \mathsf{Min}_\leq(f^{-1}[e(s)])$ for all $s \in S$. To show that $a = a^M$ satisfies (EC) assume we have an $s \in S$ and $T \subseteq S$ such that $e(s) = \bigcup\{e(t) \mid t \in T\}$. Then take any $p \in \mathsf{At}$ such that $p \in a^M(t)$ for all $t \in T$. We need to show that $p \in a^M(s)$. In the following it is convenient to use the function $e' : S \to \mathcal{P}W'$ which is defined such that $e'(s) = f^{-1}[e(s)]$ for all $s \in S$. Because $f^{-1}[\cdot]$ preserves all Boolean operations it follows from the assumption $e(s) = \bigcup\{e(t) \mid t \in T\}$ that

$$e'(s) = \bigcup\{e'(t) \mid t \in T\}. \tag{7.5}$$

That $p \in a^M(t)$ for all $t \in T$ means that for all $t \in T$ the conditional $e'(t) \mathrel{\mid\!\sim} I(p)$ is true on $\leq$. We show that for this set of conditionals $\Phi = \{e'(t) \mathrel{\mid\!\sim} I(p) \mid t \in T\}$ we have that the inference $\Phi/e'(s) \mathrel{\mid\!\sim} I(p)$ is derivable in P_∞. By Theorem 7.4.2 it then follows that $e'(s) \mathrel{\mid\!\sim} I(p)$ is true on $\leq$, which gives the required $p \in a^M(s)$.

That the inference $\Phi/e'(s) \mathrel{\mid\!\sim} I(p)$ is derivable in P_∞ follows because we can see that the two conditions in Theorem 7.4.1 are satisfied by setting $\Psi = \Phi$. If we instantiate the first condition in Theorem 7.4.1 with our inference we obtain

$$e'(s) \subseteq I(p) \cup \bigcup\{e'(t) \setminus I(p) \mid t \in T\}.$$

To see that this holds consider any $w \in e'(s)$ and distinguish cases on $w \in I(p)$. If $w \in I(p)$ then we are done because $I(p)$ is clearly contained in set on the right above. If $w \notin I(p)$ then we use the fact that $w \in e'(s)$. Because of (7.5) it follows that there is a $t \in T$ such that $w \in e'(t)$. But then also $w \in e'(t) \setminus I(p)$ and we are done. Instantiating the second condition in Theorem 7.4.1 yields that we need for all $t \in T$ that $e'(t) \cap I(p) \subseteq I(p) \cup U(\Psi)$. But this is trivially the case.

We now show that any behavior $a : S \to \mathcal{P}W$ which satisfies the exact covering condition (EC) with respect to some evidence function $e : S \to \mathcal{P}W$ is splitting plausibility interpretable with e. So assume that a and e satisfy (EC). We need to construct some splitting plausibility model $(W', f, \leq, I)$ such that $a = a^M$ for the multi-situation model $M = (W', b, I)$ where $b' : S \to \mathcal{P}W'$ is defined such that $b(s) = \mathsf{Min}_{\leq}(f^{-1}[e(s)])$ for all $s \in S$.

We start by considering the set of worlds $W'' = W \times \mathcal{P}\mathsf{At}$ consisting of all pairs (w, c) such that $w \in W$ is some world from W and $c \subseteq \mathsf{At}$ is a set of atomic sentences. For this set we have the projection $p : W'' \to W$ which is a surjective function mapping a pair $(w, c) \in W''$ to its first component w. There is a natural way to define an interpretation function $I'' : \mathsf{At} \to \mathcal{P}W''$ for W'' by setting for all $p \in \mathsf{At}$

$$I''(p) = \{(w, c) \in W'' \mid p \in c\}.$$

We also lift the evidence function $e : S \to \mathcal{P}W$ along $p : W'' \to W$ to obtain a function $e'' : S \to \mathcal{P}W''$ defined such that $e''(s) = p^{-1}[e(s)]$ for all $s \in S$.

Consider following set of conditionals over W''

$$\Phi = \{e''(t) \mathrel{\mid\!\sim} I''(q) \mid t \in S, q \in a(t)\}.$$

This set contains precisely the conditionals that we want to be true in the well-founded preorder that is used to witnesses interpretability of a. To get such an order we use the completeness of system P_∞ as stated by Theorem 7.4.3. From this theorem we obtain a well-founded poset $\leq$ on some set W' together with a function $g : W' \to W''$ such that for all $A, C \subseteq W''$ the conditional $g^{-1}[A] \mathrel{|\!\!\sim} g^{-1}[C]$ is true in $\leq$ iff $\Phi/A \mathrel{|\!\!\sim} C$ is provable in P_∞.

We then define the model $M = (W', f, \leq, I)$ such that $f(w) = p(g(w))$ for all $w \in W'$ and $I(p) = g^{-1}[I''(p)]$ for all $p \in \mathsf{At}$. This model witnesses the splitting plausibility interpretability of a with e because we can show that $a = a^M$ for the multi situation model $M = (W', b, I)$ where $b : S \to \mathcal{P}W'$ is such that $b(s) = \mathsf{Min}_\leq(g^{-1}[p^{-1}[e(s)]])$ for all $s \in S$. To show that $a = a^M$ we show that for all $s \in S$ and $p \in \mathsf{At}$

$$p \in a(s) \quad \text{iff} \quad p \in a^M(s).$$

First take any $p \in a(s)$ for some $s \in S$. We need that $p \in a^M(s)$. This is equivalent to showing that $\mathsf{Min}_\leq(g^{-1}[e''(s)]) \subseteq I(p)$ which in turn means that $g^{-1}[e''(s)] \models g^{-1}[I''(p)]$ is true on $\leq$. By the construction of $\leq$ the latter is true precisely if $\Phi/e''(s) \mathrel{|\!\!\sim} I''(p)$ is provable in P_∞. But this is trivially the case because $e''(s) \mathrel{|\!\!\sim} I''(p) \in \Phi$.

For the other inclusion assume that $p \in a^M(s)$ for some $s \in S$. We want to show that $p \in a(s)$. That $p \in a^M(s)$ means that $\mathsf{Min}_\leq(g^{-1}[p^{-1}[e(s)]]) \subseteq I(p)$, which is equivalent to saying that $g^{-1}[e''(s)] \mathrel{|\!\!\sim} g^{-1}[I''(p)]$ is true in $\leq$. By the construction of $\leq$ this amounts to the claim that $\Phi/e''(s) \mathrel{|\!\!\sim} I''(p)$ is provable in P_∞. By Theorem 7.4.1 it follows that there is some $\Psi \subseteq \Phi$ such that the two conditions of the theorem are satisfied. We show that form this we can construct a suitable covering such that $\varphi \in a(s)$ follows by an application of (EC). The first condition of Theorem 7.4.1 is that

$$e''(s) \subseteq I''(p) \cup U(\Psi).$$

Now consider any world $w \in e(s)$ and define $c = \mathsf{At} \setminus \{p\}$. Consider the pair $(w, c) \in p^{-1}[e(s)] = e''(s)$. By the inclusion above $(w, c) \in I''(p) \cup U(\Psi)$. Because $p \notin c$ we have that $(w, c) \notin I''(p)$. So it follows that $(w, c) \in U(\Psi)$. This means that there is some $t_w \in S$ and some $q_w \in a(t_w)$ such that $(w, c) \in e''(t_w) \setminus I''(q_w)$ and $e''(t_w) \mathrel{|\!\!\sim} I''(q_w) \in \Psi$. From $(w, c) \in e''(t_w)$ it follows that $w \in e(t_w)$ and from $(w, c) \notin I''(q_w)$ it follows that $q_w \notin c = \mathsf{At} \setminus \{p\}$. The latter entails that $q_w = p$. So in total we have for every $w \in e(s)$ some $t_w \in S$ such that $w \in e(t_w)$, $p \in a(t_w)$ and $e''(t_w) \mathrel{|\!\!\sim} I''(p) \in \Psi$.

We can now set $T = \{t_w \mid w \in e(s)\}$. By construction $e(s) \subseteq \bigcup\{e(t) \mid t \in T\}$. Because $p \in a(t)$ for all $t \in T$ we already almost have a covering that would allow us to conclude that $p \in a(s)$ by using the exact covering condition (EC). What remains to be shown is that the covering is exact meaning that $e(t_w) \subseteq e(s)$ for all $w \in e(s)$.

So take any $w \in e(s)$ and pick $v \in e(t_w)$. We want to show that $v \in e(s)$. Because $e''(t_w) \mathrel{|\!\!\sim} I''(p) \in \Psi$ we can use the second condition from Theorem 7.4.1 which is

$$e''(t_w) \cap I''(p) \subseteq (e''(s) \cap I(p)) \cup U(\Psi).$$

Consider the world $(v, \mathsf{At}) \in W''$. Because $v \in e(t_w)$ we have that $(v, \mathsf{At}) \in e''(t_w)$ and clearly $(v, \mathsf{At}) \in I''(p)$ so we obtain from the above inclusion that $(v, \mathsf{At}) \in (e''(s) \cap I''(p)) \cup U(\Psi)$. We now derive a contradiction from the assumption that $(v, \mathsf{At}) \in U(\Psi)$. This entails that $(v, \mathsf{At}) \in e''(s) \cap I''(p)$ and hence $v \in e(s)$. So assume for a contradiction that $(v, \mathsf{At}) \in U(\Psi)$. Then there is some $r \in S$ and $q \in \mathsf{At}$ such that $e''(r) \vdash I''(q) \in \Psi$ and $(v, \mathsf{At}) \in e''(r) \setminus I''(p)$. This would mean especially that $(v, \mathsf{At}) \notin I''(p)$ and hence by the definition of I'' we get the contradiction that $p \notin \mathsf{At}$. □

In the case where $\mathcal{V} = \mathcal{B}$ the proof of the representation result uses the same strategy as in the case of atomic sentences. I recommend that the reader first study the proof of Theorem 7.4.4 before looking at the proof of the following theorem:

7.4.5. Theorem. *A behavior $a : S \to \mathcal{PB}$ is plausibility interpretable with $e : S \to \mathcal{P}W$ iff it satisfies the plausibility covering condition.*

Proof. For the left-to-right direction assume that the behavior $a : S \to \mathcal{PB}$ is plausibility interpretable with the evidence function $e : S \to \mathcal{P}W$. So there is a splitting plausibility model $(W', f, \leq, I)$ such that $a = a^M$ for the multi-situation model $M = (W', b', I)$ where $b' : S \to \mathcal{P}W'$ is defined such that $b'(s) = \mathsf{Min}_{\leq}(f^{-1}[e(s)])$ for all $s \in S$.

We want to show that a^M satisfies the plausibility conditions. Hence assume that we have an $X \subseteq W$ with $e(s) \subseteq X$ such that all $w \in e(s)$ are potentially $\varphi \in \mathcal{M}$ relative to X and all $v \in X \setminus e(s)$ are implausible relative to X. It needs to follow that $\varphi \in a^M(s)$.

Consider now any $w \in e(s)$. Because w is potentially φ in X there exists $t_{w,1}, \ldots, t_{w,n_w} \in S$ with $w \in e(t_{w,k}) \subseteq X$ for all $k \in \{1, \ldots, n_w\}$ such that $\varphi \in \mathsf{cl}\,(a(t_{w,1}) \cup \cdots \cup a(t_{w,n_w}))$. By compactness we can reduce the latter to the claim that for each $k \in \{1, \ldots, n_w\}$ there is a $\psi_{w,k} \in a(t_{w,k})$ such that $\psi_{w,1} \wedge \cdots \wedge \psi_{w,n_w} \models \varphi$.

A similar property holds for any fixed $v \in X \setminus e(s)$. Because v is implausible in X there exists $r_{v,1}, \ldots, r_{v,m_v} \in S$ with $v \in e(r_{v,l}) \subseteq X$ for all $l \in \{1, \ldots, m_v\}$ such that $\bot \in \mathsf{cl}\,(a(r_{v,1}) \cup \cdots \cup a(r_{v,m_v}))$. By compactness we can reduce the latter to the claim that for each $l \in \{1, \ldots, m_v\}$ there is a $\chi_{v,l} \in a(r_{v,l})$ such that $\chi_{v,1} \wedge \cdots \wedge \chi_{v,m_v} \models \bot$.

In the following it is convenient to use the function $e' : S \to \mathcal{P}W'$ that we define such that $e'(s) = f^{-1}[e(s)]$ for all $s \in S$. Also set $X' = f^{-1}[X]$. Because $f^{-1}[\cdot]$ preserves inclusions we have that the inequalities that hold for X and e also hold for X' and e'. For instance we have that for all $w \in e(s)$ and $k \in \{1, \cdots, n_w\}$ it holds that $f^{-1}[\{w\}] \subseteq e'(t_{w,k}) \subseteq X'$ and similarly for the $e'(r_{v,l})$.

We now consider the following set of conditionals

$$\begin{aligned}\Phi = \{e'(t_{w,k}) \mathrel{|\!\!\sim} I(\psi_{w,k}) \mid k \in \{1, \dots, n_w\}, w \in e(s)\} \cup \\ \{e'(r_{v,l}) \mathrel{|\!\!\sim} I(\chi_{v,l}) \mid l \in \{1, \dots, m_v\}, v \in X \setminus e(s)\}.\end{aligned}$$

We proceed by showing that all conditionals in Φ are true on $\leq$ and that the inference $\Phi/e'(s) \mathrel{|\!\!\sim} I(\varphi)$ is derivable in the system P_∞. By the soundness of P_∞, as stated in Theorem 7.4.2, it then follows that $e'(s) \mathrel{|\!\!\sim} I(\varphi)$ is true in $\leq$ which means that $b'(s) = \mathsf{Min}_{\leq}(e'(s)) \subseteq I(\varphi)$ and so we get the desired $\varphi \in a^M(s)$.

The conditionals of the form $e'(t_{w,k}) \mathrel{|\!\!\sim} I(\psi_{w,k})$ for $k \in \{1, \dots, n_w\}$, and $w \in e(s)$ are true in $\leq$ because by the choice of $\psi_{w,k}$ we have that $\psi_{w,k} \in a^M(t_{w,k})$ and so $\mathsf{Min}_{\leq}(e'(t_{w,k})) \subseteq I(\psi_{w,k})$. Similarly, the conditionals of the form $e'(r_{v,l}) \mathrel{|\!\!\sim} I(\chi_{v,l})$, for $l \in \{1, \dots, m_j\}$ and $v \in X \setminus e(s)$, are true in $\leq$ because $\chi_{v,l} \in a^M(r_{v,l})$ and hence $\mathsf{Min}_{\leq}(e'(r_{v,l})) \subseteq I(\chi_{v,l})$.

It remains to show that the inference Φ / $e'(s) \mathrel{|\!\!\sim} I(\varphi)$ is derivable in P_∞. To do so we check that the two conditions given in Theorem 7.4.1 are satisfied by setting $\Psi = \Phi$. For this we first show the following inequality:

$$X' \subseteq (e'(s) \cap I(\varphi)) \cup U(\Phi), \tag{7.6}$$

where $U(\Phi)$ is determined by the definition in Theorem 7.4.1 to be

$$\begin{aligned}U(\Phi) = \{e'(t_{w,k}) \setminus I(\psi_{w,k}) \mid k \in \{1, \dots, n_w\}, w \in e(s)\} \cup \\ \{e'(r_{v,l}) \setminus I(\chi_{v,l}) \mid l \in \{1, \dots, m_v\}, v \in X \setminus e(s)\}.\end{aligned}$$

To check this consider an arbitrary $w' \in X'$. We set $w = f(w')$. Distinguish cases on whether $w' \in e'(s)$ or $w' \in X' \setminus e'(s)$.

If $w' \in e'(s)$ then we further distinguish cases on whether $w' \in I(\varphi)$. If this is the case then we are already done because then $w' \in e'(s) \cap I(\varphi)$ which is the first disjunct of the right hand side in (7.6). So assume that $w' \notin I(\varphi)$. We show that $w' \in U(\Phi)$. Because $w' \in e'(s)$ we also have that $w \in e(s)$ and hence there are the $t_{w,1}, \dots, t_{w,n_w} \in S$ with $w' \in e'(t_{w,k})$ for all $k \in \{1, \dots, n_w\}$ and there are the $\psi_{w,1}, \dots, \psi_{w,n_w}$ such that $\psi_{w,1} \wedge \dots \wedge \psi_{w,n_w} \models \varphi$. From the contraposition of the latter it follows together with $w' \notin I(\varphi)$ that there is some $k \in \{1, \dots, n_w\}$ such that $w' \notin I(\psi_{w,k})$. And so $w' \in e'(t_{w,k}) \setminus I(\psi_{w,k}) \subseteq U(\Phi)$.

In the case where $w' \notin e'(s)$ we show that $w' \in U(\Phi)$. Because $w \in X \setminus e(s)$ there are the $r_{w,1}, \dots, r_{w,m_w} \in S$ with $w' \in e'(r_{w,l})$ for all $l \in \{1, \dots, m_w\}$ and there are the $\chi_{w,1}, \dots, \chi_{w,m_w}$ such that $\chi_{w,1} \wedge \dots \wedge \chi_{w,m_w} \models \bot$. From the latter it follows by contraposition that $\neg\chi_{w,1} \vee \dots \vee \neg\chi_{w,m_w}$ is a tautology. Hence there is some $l \in \{1, \dots, m_w\}$ such that $w' \notin I(\chi_{w,l})$. It follows that $w' \in e'(r_{w,l}) \setminus I(\chi_{w,l}) \subseteq U(\Phi)$.

We now verify the first condition of Theorem 7.4.1. Instantiating the condition yields

$$e'(s) \subseteq I(\varphi) \cup U(\Phi).$$

This follows immediately from (7.6) because $e'(s) \subseteq X'$.

For the second condition in Theorem 7.4.1 we need to check that for all $w \in e(s)$ and $k \in \{1, \dots, n_w\}$

$$e'(t_{w,k}) \cap I(\psi_{w,k}) \subseteq (e'(s) \cap I(\varphi)) \cup U(\Phi),$$

and that for all $v \in X \setminus e(s)$ and $l \in \{1, \dots, m_v\}$

$$e'(r_{v,l}) \cap I(\psi_{v,l}) \subseteq (e'(s) \cap I(\varphi)) \cup U(\Phi).$$

These also follow immediately from (7.6) because $e'(t_{w,k}) \subseteq X'$ and $e'(r_{v,l}) \subseteq X'$. This finishes the proof of the left-to-right direction of the theorem.

We now show that any behavior $a : S \to \mathcal{P}W$ which satisfies the plausibility covering condition with respect to some evidence function $e : S \to \mathcal{P}W$ is splitting plausibility interpretable with e. So assume that a and e satisfy the plausibility covering condition. We need to construct an interpreting splitting plausibility model.

We start by considering the set of worlds $W'' = W \times \mathsf{MC}(\mathcal{B})$ which consists of all pairs (w, Σ) where $w \in W$ is some world form W and $\Sigma \in \mathsf{MC}(\mathcal{B})$ is a maximal consistent set of propositional formulas. For this set we have the projection $p : W'' \to W$ which maps a pair $(w, \Sigma) \in W''$ to its first component w. The interpretation function $I'' : \mathsf{At} \to \mathcal{P}W''$ is defined in the natural way by setting for all $p \in \mathsf{At}$

$$I''(p) = \{(w, \Sigma) \in W'' \mid p \in \Sigma\}.$$

This definition clearly extends I'' to all formulas in the sense that for all $(w, \Sigma) \in W''$ and formulas $\varphi \in \mathcal{B}$

$$(w, \Sigma) \in I''(\varphi) \quad \text{iff} \quad \varphi \in \Sigma. \tag{7.7}$$

We also lift the evidence function $e : S \to \mathcal{P}W$ along $p : W'' \to W$ to obtain a function $e'' : S \to \mathcal{P}W''$ defined such that $e''(s) = p^{-1}[e(s)]$ for all $s \in S$.

Now, consider the following set of conditionals over W''

$$\Phi = \{e''(t) \mathrel{|\!\sim} I''(\psi) \mid t \in S, \psi \in a(s)\}.$$

This set contains all the conditionals that we want to be true in the well-founded preorder that witnesses interpretability of a. To obtain such an order we use the completeness of system P_∞ as stated in Theorem 7.4.3. From this theorem we obtain a well-founded poset $\leq$ on some set W' together with a function $g : W' \to W''$ such that for all $A, C \subseteq W''$ the conditional $g^{-1}[A] \mathrel{|\!\sim} g^{-1}[C]$ is true on $\leq$ iff $\Phi \mathbin{/} A \mathrel{|\!\sim} C$ is provable in P_∞.

We then define the splitting plausibility model $M = (W', f, \leq, I)$ such that $f(w) = p(g(w))$ for all $w \in W'$ and $I(p) = g^{-1}[I''(p)]$ for all $p \in \mathsf{At}$. This model witnesses the splitting plausibility interpretability of a with e because we can show that $a = a^M$ for the multi-situation model $M = (W', b, I)$ where $b : S \to \mathcal{P}W'$ is defined such that $b(s) = \mathsf{Min}_{\leq}(f^{-1}[e(s)])$ for all $s \in S$.

First take any $\varphi \in a(s)$ for some $s \in S$. We need that $\varphi \in a^M(s)$. This is equivalent to showing that $\mathsf{Min}_{\leq}(g^{-1}[e''(s)]) \subseteq I(\varphi)$ which just means that $g^{-1}[e''(s)] \mathrel{|\!\!\sim} g^{-1}[I''(\varphi)]$ is true on $\leq$. By the construction of $\leq$ the latter holds precisely if $\Phi/e''(s) \mathrel{|\!\!\sim} I''(\varphi)$ is provable in P_∞. But as observed above, right after Theorem 7.4.1, this follows immediately from $e''(s) \mathrel{|\!\!\sim} I''(\varphi) \in \Phi$.

For the other inclusion assume that $\varphi \in a^M(s)$ for some $s \in S$. We want to show that $\varphi \in a(s)$.

We first treat the trivial case where φ is a propositional tautology. In this case we can apply the plausibility covering condition with $X = e(s)$. The antecedent of the condition is satisfied because every $w \in e(s)$ is potentially φ in X, since φ is a tautology and hence $\varphi \in \mathsf{cl}\,(a(s))$ and $w \in e(s) \subseteq X$. The desired $\varphi \in a(s)$ follows as the consequent of the plausibility covering condition.

Now consider the case where φ is not a tautology. From unfolding the definition of a^M it follows from our assumption $\varphi \in a^M(s)$ that $\mathsf{Min}_{\leq}(g^{-1}[e''(s)]) \subseteq I(\varphi)$. This is the same as saying that $g^{-1}[e''(s)] \mathrel{|\!\!\sim} g^{-1}[I''(\varphi)]$ is true on $\leq$ which by construction of $\leq$ is the same as the claim that $\Phi/e''(s) \mathrel{|\!\!\sim} I''(\varphi)$ is provable in P_∞.

By Theorem 7.4.1 it follows that there is some $\Psi \subseteq \Phi$ that satisfies the two conditions of the theorem with respect to the inference $\Phi/e''(s) \mathrel{|\!\!\sim} I''(\varphi)$. Define

$$X = \bigcup\{e(t) \mid e''(t) \mathrel{|\!\!\sim} I''(\psi) \in \Psi \text{ for some } t \in S \text{ and } \psi \in \mathcal{B}\}.$$

The proofs proceeds by showing that this X satisfies the constraints of the plausibility covering condition, from which it then follows that $\varphi \in a(s)$ as required.

We first show that $e(s) \subseteq X$. For this we need the first condition of Theorem 7.4.1, which is that

$$e''(s) \subseteq I''(\varphi) \cup U(\Psi). \tag{7.8}$$

Take any world $w \in e(s)$. We want to show that $w \in X$, which amounts to showing that there is some $t \in S$ and a $\psi \in \mathcal{B}$ such that $e''(t) \mathrel{|\!\!\sim} I''(\psi) \in \Psi$ and $w \in e(t)$. Because we assume that φ is not a tautology we have that $\varphi \notin \mathsf{cl}\,(\emptyset)$ and so by Lindenbaum's Lemma there exists a $\Sigma \in \mathsf{MC}(\mathcal{B})$ such that $\varphi \notin \Sigma$. Now consider the pair $(w, \Sigma) \in p^{-1}[e(s)] = e''(s)$. By the inclusion (7.8) above $(w, \Sigma) \in I''(\varphi) \cup U(\Psi)$. Because $\varphi \notin \Sigma$ we have that $(w, \Sigma) \notin I''(\varphi)$. So it follows that $(w, \Sigma) \in U(\Psi) = \bigcup\{e''(t) \setminus I''(\psi) \mid e''(t) \mathrel{|\!\!\sim} I(\psi) \in \Psi\}$. This means that there is some $t \in S$ and a $\psi \in \mathcal{B}$ such that $e''(t) \mathrel{|\!\!\sim} I''(\psi) \in \Psi$ and $(w, \Sigma) \in e''(t) \setminus I''(\psi)$. From the latter it follows that $(w, \Sigma) \in e''(t) = p^{-1}[e(t)]$ and hence $w \in e(t)$.

It remains to be shown that all $w \in e(s)$ are potentially φ in X and that all $v \in X \setminus e(s)$ are implausible in X. In both cases we use the second condition that

Theorem 7.4.1, which in the present case becomes that for all $t \in S$ and $\varphi \in \mathcal{B}$ such that $e''(t) \mathrel{\mid\!\sim} I''(\psi) \in \Psi$

$$e''(t) \cap I''(\psi) \subseteq (e''(s) \cap I''(\varphi)) \cup U(\Psi). \tag{7.9}$$

We now show that any $w \in e(s)$ is potentially φ in X. For this we prove that

$$\varphi \in \mathsf{cl}\left(\bigcup\{a(t) \mid t \in T, w \in e(t), e(t) \subseteq X\}\right).$$

This is sufficient to show that w is potentially φ in X because we can use compactness to reduce the infinite union to a finite one.

So assume for a contradiction that $\varphi \notin \mathsf{cl}\left(\bigcup\{a(t) \mid t \in T, w \in e(t) \subseteq X\}\right)$. Then it follows by Lindenbaum's Lemma that there is some $\Sigma \in \mathsf{MC}(\mathcal{B})$ such that $\varphi \notin \Sigma$ and $a(t') \subseteq \Sigma$ for all $t \in T$ with $w \in e(t) \subseteq X$. Now consider the world $(w, \Sigma) \in W''$. Because $w \in e(s)$ and we have already shown that $e(s) \subseteq X$ there is some $t \in S$ and a $\psi \in \mathcal{B}$ such that $e''(t) \in I(\psi) \in \Psi$ and $w \in e(t) \subseteq X$. From $e''(t) \in I(\psi) \in \Psi$ and $\Psi \subseteq \Phi$ it also follows that $\psi \in a(s)$. We now instantiate the inclusion (7.9) with this $e''(t) \in I(\psi) \in \Psi$. We have that $(w, \Sigma) \in e''(t) = p^{-1}[e(t)]$ because $w \in e(t)$. Also $(w, \Sigma) \in I''(\psi)$ because $\psi \in a(t)$ and $a(t) \subseteq \Sigma$, where the latter follows from the construction of Σ because $w \in e(t) \subseteq X$. So it follows that $(w, \Sigma) \in e''(t) \cap I''(\psi)$ and hence by (7.9) we get $(w, \Sigma) \in (e''(s) \cap I''(\varphi)) \cup U(\Psi)$. But $(w, \Sigma) \notin I''(\varphi)$ because $\varphi \notin \Sigma$. Hence $(w, \Sigma) \in U(\Psi) = \bigcup\{e''(t') \setminus I''(\psi') \mid e''(t') \mathrel{\mid\!\sim} I(\psi') \in \Psi\}$. This means that there is some $e''(t') \mathrel{\mid\!\sim} I(\psi') \in \Psi$ such that $w \in e(t')$ and $\psi' \notin \Sigma$. From $e''(t') \mathrel{\mid\!\sim} I(\psi') \in \Psi$ and $\Psi \subseteq \Phi$ we obtain that $\psi' \in a(t')$. Moreover it follows from the construction of Σ that $a(t') \subseteq \Sigma$ because $w \in e(t') \subseteq X$, where the latter inclusion holds by the definition of X. Chaining these together yields $\psi' \in a(t') \subseteq \Sigma$, which contradicts the already established $\psi' \notin \Sigma$.

Now consider any $v \in X \setminus e(s)$. We need to show that v is implausible in X. The reasoning is similar to the previous two paragraphs. So we show that

$$\bot \in \mathsf{cl}\left(\bigcup\{a(t) \mid t \in T, v \in e(t), e(t) \subseteq X\}\right).$$

By compactness this suffices to establish that v is implausible in X.

Assume for a contradiction that $\bot \notin \mathsf{cl}\left(\bigcup\{a(t) \mid t \in T, v \in e(t) \subseteq X\}\right)$. Then it follows by Lindenbaum's Lemma that there is some $\Sigma \in \mathsf{MC}(\mathcal{B})$ such that and $a(t') \subseteq \Sigma$ for all $t \in T$ with $v \in e(t) \subseteq X$. Now consider the world $(v, \Sigma) \in W''$. Because $v \in X$ there is some $t \in S$ and a $\psi \in \mathcal{B}$ such that $e''(t) \in I(\psi) \in \Psi$ and $v \in e(t) \subseteq X$. From $e''(t) \in I(\psi) \in \Psi$ and $\Psi \subseteq \Phi$ it also follows that $\psi \in a(s)$. We now instantiate the inclusion (7.9) with this $e''(t) \in I(\psi) \in \Psi$. We have that $(v, \Sigma) \in e''(t) = p^{-1}[e(t)]$ because $v \in e(t)$. Also $(v, \Sigma) \in I''(\psi)$ because $\psi \in a(t)$ and $a(t) \subseteq \Sigma$, where the latter follows from the construction of Σ because $v \in e(t) \subseteq X$. So it follows that $(v, \Sigma) \in e''(t) \cap I''(\psi)$ and hence by (7.9) we get $(v, \Sigma) \in (e''(s) \cap I''(\varphi)) \cup U(\Psi)$. But $(v, \Sigma) \notin e''(s)$ because $v \notin e(s)$. Hence

$(v, \Sigma) \in U(\Psi) = \bigcup\{e''(t') \setminus I''(\psi') \mid e''(t') \mathrel{\vdash\!\!\sim} I(\psi') \in \Psi\}$. This means that there is some $e''(t') \mathrel{\vdash\!\!\sim} I(\psi') \in \Psi$ such that $v \in e(t')$ and $\psi' \notin \Sigma$. From $e''(t') \mathrel{\vdash\!\!\sim} I(\psi') \in \Psi$ and $\Psi \subseteq \Phi$ we obtain that $\psi' \in a(t')$. Moreover it follows from the construction of Σ that $a(t') \subseteq \Sigma$ because $v \in e(t') \subseteq X$, where the latter inclusion holds by the definition of X. Chaining these together yields $\psi' \in a(t') \subseteq \Sigma$, which contradicts the already established $\psi' \notin \Sigma$. □

7.5 Supervaluations

This sections contains proofs for Chapter 5. The main result is that supervaluation interpretability with a belief function is the same as splitting interpretability. I then also give a counterexample which shows that in the setting of plausibility orders this is no longer the case.

7.5.1. Theorem. *A behavior $a : S \to \mathcal{PB}$ is supervaluation interpretable with a belief function $b : S \to \mathcal{P}W$ iff a is splitting interpretable with b.*

Proof. We first show the left-to-right direction. So assume that the behavior $a : S \to \mathcal{PB}$ is supervaluation interpretable with a belief function $b : S \to \mathcal{P}W$. So there exists a supervaluation model $M = (W, b, \mathcal{I})$ such that $a = a^M$. To show that a is splitting interpretable with b we have to find a splitting interpretation model (W', f, I') such that $a^M = a^{M'}$ for the multi situation model $M' = (W', b', \mathcal{I})$, where $b' : S \to \mathcal{P}W'$ is such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$.

We define the domain W' of M' to be the product $\mathcal{I} \times W$. The splitting function $f : W' \to W$ is the projection to the second component, meaning that $f(I, w) = w$ for all $(I, w) \in \mathcal{I} \times W$. This function is clearly surjective because the set $\mathcal{I}$ is a supervaluation and hence not empty. We choose the interpretation function $I' : \mathsf{At} \to \mathcal{P}W'$ such that $I'(p) = \{(I, w) \mid w \in I(p)\}$ for all $p \in \mathsf{At}$. By an induction on the complexity of the propositional formula $\varphi \in \mathcal{B}$ one can show that this definition makes it the case that for all $I \in \mathcal{I}$

$$w \in I(\varphi) \quad \text{iff} \quad (I, w) \in I'(\varphi). \tag{7.10}$$

To show that $a^M = a^{M'}$ we need that for all $s \in S$ and $\varphi \in \mathcal{B}$ it holds that

$$b(s) \subseteq I(\varphi) \text{ for all } I \in \mathcal{I} \quad \text{iff} \quad b'(s) \subseteq I'(\varphi).$$

So first assume the left side and pick any $(I, w) \in b'(s)$. We want to show that $(I, w) \in I'(\varphi)$. From $(I, w) \in b'(s) = f^{-1}[b(s)]$ it follows that $w \in b(s)$. By the assumption that $b(s) \subseteq I(\varphi)$ it follows that $w \in I(\varphi)$ which by (7.10) implies that $(I, w) \in I'(\varphi)$. For the other direction assume that $b'(s) \subseteq I'(\varphi)$. We want to show that $b(s) \subseteq I(\varphi)$ for all $I \in \mathcal{I}$. So fix any $I \in \mathcal{I}$ and consider some

$w \in b(s)$. We want to have that $w \in I(\varphi)$. Because $w \in b(s)$ we have that $(w, I) \in f^{-1}[b(s)] = b'(s)$. Hence $(I, w) \in I'(\varphi)$ and it follows by (7.10) that $w \in I(\varphi)$.

For the other direction assume that we have a linguistic behavior a is splitting interpretable with $b : S \to \mathcal{P}W$. Hence there is some splitting interpretation model (W', f, I) over W such that $a = a^{M'}$ for the multi-situation model $M' = (W', b', I)$ where $b' : S \to \mathcal{P}W$ is defined such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$. We need to find a supervaluation $\mathcal{I}$ over W show that $a = a^M$ for the supervaluation model $M = (W, b, \mathcal{I})$.

To obtain the supervaluation $\mathcal{I}$ define the index set J as the supremum of the cardinalities of all the sets $f^{-1}[\{w\}]$ for some $w \in W$. Because f is surjective we have $f^{-1}[\{w\}]$ is not empty for all $w \in W$. Together with the choice of J this implies that every $w \in W$ there exists a surjective function $g_w : J \to f^{-1}[\{w\}]$. We can now define $\mathcal{I} = \{I_j \mid j \in J\}$ to be a set of J-many interpretation functions such that for any $j \in J$ the interpretation $I_j : \mathsf{At} \to \mathcal{P}W$ is defined such that for every $p \in \mathsf{At}$

$$I_j(p) = \{w \in W \mid g_w(j) \in I'(p)\}.$$

We can assume without loss of generality that the set $\mathcal{I}$ is not empty and hence a supervaluation. If it was empty then J would also be empty. Because f is surjective this would entail that W is empty. But then $a^M(s)$ is inconsistent for every $s \in S$ and hence a^M is interpretable by the supervaluation model $(W, b, \mathcal{I})$ where $\mathcal{I} = \{I\}$ and $I : \mathsf{At} \to \mathcal{P}W$ is such that $I(p) = \emptyset = W$ for all $p \in \mathsf{At}$.

By an induction on the complexity of the formula $\varphi \in \mathcal{B}$ on can prove that for all $j \in J$.

$$w \in I_j(\varphi) \quad \text{iff} \quad g_w(j) \in I'(\varphi). \tag{7.11}$$

To show that $a^M = a^{M'}$ we need that for all $s \in S$ and $\varphi \in \mathcal{B}$ it holds that

$$b(s) \subseteq I_j(\varphi) \text{ for all } j \in J \quad \text{iff} \quad b'(s) \subseteq I'(\varphi).$$

So first assume the left side and pick any $w' \in b'(s)$. We want to show that $w' \in I'(\varphi)$. From $w' \in b'(s) = f^{-1}[b(s)]$ it follows that $f(w') \in b(s)$. By the assumption this means that for every $j \in J$ we have that $f(w') \in I_j(\varphi)$ which by (7.11) is equivalent to $g_{f(w')}(j) \in I'(\varphi)$. So to get that $w' \in I'(\varphi)$ it suffices to show that $w' = g_{f(w')}(j)$ for some $j \in J$. But such a $j \in J$ exists because $g_{f(w')} : J \to f^{-1}[\{w'\}]$ is surjective and clearly $w' \in f^{-1}[\{f(w')\}]$.

For the other direction assume that $b'(s) \subseteq I'(\varphi)$. We want to show that $b(s) \subseteq I_j(\varphi)$ for all $j \in J$. So fix any $j \in J$ and consider some $w \in b(s)$. We want to have that $w \in I_j(\varphi)$. By the definition of $g_w : J \to f^{-1}[\{w\}]$ we get that $g_w(j) \in f^{-1}[\{w\}]$. Because $w \in b(s)$ this implies that $g_w(j) \in f^{-1}[b(s)] = b'(s)$. With the assumption $b'(s) \subseteq I'(\varphi)$ we obtain that $g_w(j) \in I'(\varphi)$. From this it follows by (7.11) that $w \in I_j(\varphi)$. □

I conclude this section with an example of a behavior that is splitting plausibility interpretable but is not supervaluation plausibility interpretable for a very weak notion of supervaluation interpretability. This is the technical example that I refer to at the end of Section 5.3.

We first need to define supervaluation plausibility interpretability. The interpreting models are of the following form:

7.5.2. Definition. A *supervaluation plausibility model over* W is defined to be a tuple $(W', f, \leq, \mathcal{I})$ such that $f : W' \to W$ is some function, $\leq \subseteq W' \times W'$ is a well-founded preorder on W' and $\mathcal{I}$ is a supervaluation over W.

The function $f : W' \to W$ in this definition gives us additional freedom in defining the plausibility order. The elements of the order need not be worlds in W, instead they are just labeled by worlds in W. This means that the same world in W can occur at difference places in the plausibility order. To get an idea of why it might be adequate to allow this freedom in defining the plausibility order the reader is referred to Remark 5 in (Marti and Pinosio 2016).

It is crucial that in the above definition the supervaluation $\mathcal{I}$ is defined over W and not over W'. So it is only the doxastic part that profits from the function f and there is no splitting of worlds with respect to the semantic facts that are captured by the supervaluation.

If the reader does not like the function $f : W' \to W$ in the definition of supervaluation plausibility models she or he might as well just think of the special case where $W' = W$ and f is the identity function. Because this yields a more restrictive notion of supervaluation plausibility interpretability the example below, of a behavior that is not supervaluation plausibility interpretable, still applies.

Next, we define supervaluation plausibility interpretability.

7.5.3. Definition. A supervaluation plausibility model $(W', f, \leq, \mathcal{I})$ over W *interprets a linguistic behavior* $a : S \to \mathcal{PV}$ *with an evidence function* $e : S \to \mathcal{P}W$ if $a = a^M$ for the supervaluation model $M = (W, b, \mathcal{I})$ where $b : S \to \mathcal{P}W'$ is defined such that for all $s \in S$

$$b(s) = f[\mathsf{Min}_{\leq}(f^{-1}[e(s)])] = \{f(w') \in W \mid w' \in \mathsf{Min}_{\leq}(f^{-1}[e(s)])\}.$$

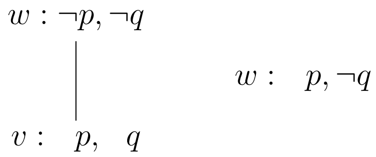
A behavior $a : S \to \mathcal{PV}$ is *supervaluation plausibility interpretable with an evidence function* $e : S \to \mathcal{P}W$ if there is some splitting plausibility model that interprets a with e.

Again, the reader might be puzzled by the choice of the belief function b in this definition. To see that the definition is not completely arbitrary one might check that for all $P \subseteq W$ it holds that $b(s) \subseteq P$ iff $\mathsf{Min}_{\leq}(f^{-1}[e(s)]) \subseteq f^{-1}[P]$. In the case where $W' = W$ and f is the identity function it just amounts to the usual $b(s) = \mathsf{Min}_{\leq}(e(s))$.

We can now give an example of a behavior that is splitting plausibility interpretable with some evidence function but not supervaluation plausibility interpretable with the same evidence function.

7.5.4. Example. Take the domain $W = \{w, v\}$ and let $S = \{s_1, s_2, s_3\}$. Consider the evidence function $e : S \to \mathcal{P}W$ such that $e(s_1) = \{w, v\}$, $e(s_2) = \{w\}$ and $e(s_3) = \{v\}$. Assume that $\mathsf{At} = \{p, q\}$ and that $\mathcal{V} = \mathcal{B}$ is the set of all propositional formulas over At. The behavior $a : S \to \mathcal{PM}$ with $a(s_1) = \mathsf{cl}\,(\{p\})$, $a(s_2) = \mathsf{cl}\,(\{\neg q\})$ and $a(s_3) = \mathsf{cl}\,(\{p, q\})$ is splitting plausibility interpretable with e but it is not supervaluation interpretable with e.

That a is splitting plausibility interpretable with e is witnessed by the splitting plausibility model $(W', f, \leq, I)$ such that $W' = \{w, w', v\}$, $f(w) = f(w') = w$, $f(v) = v$, $\leq\, = \{(w, w), (w', w'), (v, v), (v, w)\}$, $I(p) = \{v, w'\}$ and $I(q) = \{v\}$. This model is suggested by the following picture:



We can derive a contradiction from the assumption that a is supervaluation plausibility interpretable with e. In that case there would be a supervaluation plausibility model $(W', f, \leq, \mathcal{I})$ over W such that a is the behavior generated by the supervaluation model $(W, b, \mathcal{I})$, where $b : S \to \mathcal{P}W$ is as in Definition 7.5.3. Because both $a(s_2)$ and $a(s_3)$ are consistent it must be the case that $b(s_2)$ and $b(s_3)$ are not empty. Moreover, one can see from the definition of b that $b(s) \subseteq e(s)$ for all $s \in S$. So it follows that $b(s_2) = e(s_2) = \{w\}$ and $b(s_3) = e(s_3) = \{v\}$. Because $p \notin a(s_2)$ there must be some $I \in \mathcal{I}$ such that $w \notin I(p)$. It follows that $w \notin b(s_3)$ because otherwise we had $p \notin a(s_3) = \mathsf{cl}\,(\{p, q\})$. So $b(s_3)$ must be either empty or equal to $\{v\} = b(s_1)$. But $b(s_3)$ can not be empty because $a(s_3)$ is consistent and it can not be equal to $b(s_1)$ because otherwise we had that $a(s_3) = a(s_1)$.

A similar example works in the case where $\mathcal{V} = \mathsf{At}$. One can use the behavior $a : S \to \mathcal{P}\mathsf{At}$ such that $a(s_1) = \{p\}$, $a(s_2) = \emptyset$ and $a(s_3) = \{p, q\}$. An additional atomic sentence r is needed such that whenever in the above argument we use the consistency of $a(s)$ to show that $b(s)$ is not empty we would use that $r \notin a(s)$ instead of consistency. The model that witnesses the splitting plausibility interpretability has to be adapted such that r is false at all worlds.

The example seems to rely crucially on the fact that the plausibility order $\leq$ is not complete, that is, it does not satisfy that for all worlds w and v at least one of $w \leq v$ or $v \leq w$ holds. I conjecture that with the constraint that $\leq$ is complete it follows by a similar construction as in Theorem 7.5.1 that supervaluation and splitting plausibility interpretability are the same.

7.6 Modalities

In this section I discuss the results for the setting from Chapter 6 where the language of the subject contains a modal operator. This section presupposes knowledge of basic results in modal logic that can be found for instance in (Blackburn, de Rijke, and Venema 2002). I use the terminology of this book.

We start with a simple observation that is needed in Section 6.2.

7.6.1. Proposition. *If a linguistic behavior $a : S \to \mathcal{PM}$ is supervaluation interpretable with some belief function $b : S \to \mathcal{P}W$ then it satisfies that necessity is a priori.*

Proof. That $a : S \to \mathcal{PM}$ is supervaluation interpretable with $b : S \to \mathcal{P}W$ means by definition that there is some supervaluation $\mathcal{I}$ over W that $a = a^M$ for the supervaluation model $M = (W, b, \mathcal{I})$. So we need to check that a^M satisfies that necessity is a priori. For this take some $s \in S$ such that $a^M(s)$ is consistent and pick some $\Box\varphi \in a^M(s)$. We have to show that $\Box\varphi \in a^M(s')$ for any arbitrarily chosen $s' \in S$. By definition of a^M this means that we need to show that $b(s') \subseteq I(\Box\varphi)$ for all $I \in \mathcal{I}$.

Fix an $I \in \mathcal{I}$. Now first observe that because $a^M(s)$ is consistent there is a $v \in b(s)$. Since $\Box\varphi \in a^M(s)$ it needs to be the case that $v \in I(\Box\varphi)$. From the semantic clause for the modality relative to an interpretation function one can see that either $I(\Box\varphi) = \emptyset$ or $I(\Box\varphi) = W$. Because $v \in I(\Box\varphi)$ it follows that the latter is the case and hence $b(s') \subseteq I(\Box\varphi)$. □

The remaining parts of this section prove the representation results from Sections 6.3 and 6.4. I first introduce a class of models in which the modal operator is interpreted on the relation of a Kripke frame. The representation results from Sections 6.3 and 6.4 follow then from a more general representation result for this class of models.

7.6.2. Definition. A *relational model* (W, R, U, b, I) consists of a domain W, relation $R \subseteq W \times W$ on W, a set of worlds $U \subseteq W$, a belief function $b : S \to \mathcal{P}U$ and an interpretation function $I : \mathsf{At} \to \mathcal{P}W$.

The *linguistic behavior* $a^M : S \to \mathcal{PV}$ *generated* by a relational model $M = (W, R, U, b, I)$ is defined such that for all situations $s \in S$

$$a^M(s) = \{\varphi \in \mathcal{V} \mid b(s) \subseteq I(\varphi)\},$$

where $I : \mathsf{At} \to \mathcal{P}W$ extends to $\mathcal{M}$ with the usual semantic clause on Kripke frames, that is,

$$I(\Box\varphi) = \{w \in W \mid v \in I(\varphi) \text{ for all } v \in W \text{ with } wRv\}.$$

We can define a notion of interpretability for relational models that is parametric on the type of frame that an interpreting model is based on.

7.6.3. Definition. A *splitting relational model over W* is a tuple (W', R, U, f, I) such that (U, f, I) is a splitting interpretation model over W, $R \subseteq W' \times W'$ is some relation over W' and $U \subseteq W'$.

A splitting relational model (W', R, U, f, I) over W *interprets a behavior $a : S \to \mathcal{PM}$ with a belief function $b : S \to \mathcal{P}W$* if $a = a^M$ for the relational model (W', R, b', I), where $b' : S \to \mathcal{P}W'$ is defined such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$.

Let $\mathcal{F}$ be some class of Kripke frames. A linguistic behavior $a : S \to \mathcal{PM}$ is *splitting $\mathcal{F}$-interpretable with a belief function $b : S \to \mathcal{P}W$* if there exists some splitting relational model (W', R, f, U, I) that interprets a with b and such that (W', R) is a frame in $\mathcal{F}$.

I now first give a representation theorem for $\mathcal{F}$-interpretability, where $\mathcal{F}$ is some class of Kripke frames. For the proof I need the following notions:

7.6.4. Definition. For every Kripke model (W, R, I) and world $w \in W$ define the *theory of w* as the set $\mathsf{Th}(w) \subseteq \mathcal{M}$ of all formulas that are true at W, that is,

$$\mathsf{Th}(w) = \{\varphi \in \mathcal{M} \mid \varphi \in I(\varphi)\}.$$

For every class of frames $\mathcal{F}$ define the set $\mathsf{Th}(\mathcal{F}) \subseteq \mathcal{PM}$ as the set of all theories of worlds that belong to a model that are based on a frame in $\mathcal{F}$. In symbols this amounts to setting

$$\mathsf{Th}(\mathcal{F}) = \{\mathsf{Th}(w) \subseteq \mathcal{M} \mid w \in W \text{ for some Kripke model } (W, R, I), \\ \text{such that } (W, R) \text{ is a frame in } \mathcal{F}\}.$$

The next theorem characterizes $\mathcal{F}$-interpretability, for some class of frames $\mathcal{F}$, with conditions similar to the conjunctive covering and consistency conditions for local consequence on $\mathcal{F}$. They however involve infinite intersections because in general we can not assume that local consequence on $\mathcal{F}$ is compact.

7.6.5. Theorem. *Let $\mathcal{F}$ be a class of frames that is closed under coproducts. A linguistic behavior $a : S \to \mathcal{PM}$ is splitting $\mathcal{F}$-interpretable with a belief function $b : S \to \mathcal{P}W$ iff a satisfies the following infinitary conjunctive covering condition that for all $s \in S$ and $\mathcal{T} \subseteq \mathcal{P}S$*

$$b(s) \subseteq \bigcup_{T \in \mathcal{T}} \bigcap_{t \in T} b(t) \textit{ implies } \bigcap_{T \in \mathcal{T}} \mathsf{cl}\left(\bigcup_{t \in T} a(t)\right) \subseteq a(s), \tag{ICC}$$

and the following infinitary conjunctive consistency condition that for all $T \subseteq S$

$$\textit{If } \bigcap_{t \in T} b(t) \neq \emptyset \textit{ then } \mathsf{cl}\left(\bigcup_{t \in T} a(t)\right) \textit{ is consistent.} \tag{ICCon}$$

The notion of logical consequence that is used in these conditions is local consequence on the class of frames $\mathcal{F}$.

Proof. We first show the left-to-right direction.

So assume that $a : S \to \mathcal{P}\mathcal{M}$ is splitting $\mathcal{F}$-interpretable with $b : S \to \mathcal{P}W$. We need to check that then the infinitary conjunctive covering and consistency conditions hold. That a is splitting $\mathcal{F}$-interpretable means that $a = a^M$ for some relational model $M = (W', R, U, b', I)$ with $b'(s) = f^{-1}[c(s)]$ for some function $f : W' \to W$.

So assume that we are given $s \in S$ and $\mathcal{T} \subseteq \mathcal{P}S$ such that the antecedent of the infinitary conjunctive covering condition is satisfied. Moreover assume we have a $\varphi \in \mathcal{M}$ such that $\varphi \in \mathsf{cl}\left(\bigcup_{t \in T} a^M(t)\right)$ for all $T \in \mathcal{T}$. We want to show that then $\varphi \in a^M(s)$.

For this we need to show that $w \in I(\varphi)$ for any $w \in b'(s) \subseteq U$. By the fact that f^{-1} preserves intersections, unions and inclusions it follows from the antecedent of the conjunctive covering condition that $b'(s) \subseteq \bigcup_{T \in \mathcal{T}} \left(\bigcap_{t \in T} b'(t)\right)$. So there is some $T \in \mathcal{T}$ such that $w \in b'(t)$ for all $t \in T$. Because $\varphi \in \mathsf{cl}\left(\bigcup_{t \in T} a^M(t)\right)$ there is a $\Sigma \subseteq \bigcup_{t \in T} a^M(t)$ such that $\Sigma \models \varphi$ is a local consequence on $\mathcal{F}$. By the definition of a^M it follows that $w \in I(\psi)$ for all $\psi \in \Sigma$ and hence $w \in I(\varphi)$ because $\models$ is the local consequence relation for the class of frames $\mathcal{F}$ on which M is based.

To show that a^M satisfies the infinitary conjunctive consistency condition we need to take an arbitrary $T \subseteq S$ such that $\bigcap_{t \in T} b(t) \neq \emptyset$ and check that $\mathsf{cl}\left(\bigcup_{t \in T} a^M(t)\right)$ is consistent. Because $\bigcap_{t \in T} b(t) \neq \emptyset$ there is some $w \in W$ such that $w \in b(t)$ for all $t \in T$. By the subjectivity of f there is then a $w' \in W'$ such that $f(w') = w$. Because $b'(t) = f^{-1}[b(t)]$ we then also have that $w' \in b'(t)$ for every $t \in T$.

Now assume for a contradiction that $\bot \in \mathsf{cl}\left(\bigcup_{t \in T} a^M(t)\right)$. This would entail that there is a $\Sigma \subseteq \bigcup_{t \in T} a^M(t)$ such that $\Sigma \models \bot$ is a local consequence on $\mathcal{F}$. Because $w' \in b'(t)$ for all $t \in T$ we have by the definition of a^M that $w' \in I(\psi)$ for all $\psi \in \Sigma$. By the definition of local consequence it would follow that $w' \in I(\bot)$ which is impossible.

For the right-to-left direction of the first claim assume that a satisfies the two conditions. We have to construct a splitting relational model (W', R, U, f, I) such that (W', R) is in $\mathcal{F}$ and $a = a^M$ for the relational model $M = (W', R, U, b', I)$ where $b' : S \to \mathcal{P}U$ is such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$.

To define this splitting relational model first consider the following index set $D \subseteq W \times \mathsf{Th}(\mathcal{F})$:

$$D = \{(w, \Sigma) \in W \times \mathsf{Th}(\mathcal{F}) \mid a(s) \subseteq \Sigma \text{ for all } s \in S \text{ with } w \in b(s)\}.$$

By the definition of $\mathsf{Th}(\mathcal{F})$ we can pick for every $d = (w, \Sigma) \in D$ a world w_d in some Kripke model M_d based on $\mathcal{F}$ such that $\Sigma = \mathsf{Th}(w_d)$.

Let the Kripke model $(W', R, I) = \coprod_{d \in D} M_d$ be the coproduct of all the Kripke models M_d for some $d \in D$. Intuitively, this model results by placing all the individual models M_d next to each other without adding any additional relations.

We also write w_d for the copy of w_d in the coproduct (W', R, I). Because the construction of the coproduct preserves the truth of modal formulas we have that $\Sigma = \mathsf{Th}(w_d)$ for all $d \in D$, where by w_d we now mean the world in the coproduct (W', R, I).

Because $\mathcal{F}$ is assumed to be closed under coproducts it follows that (W', R) is a frame in $\mathcal{F}$.

We define $U \subseteq W'$ such that $U = \{w_d \mid d \in D\}$ and $f : U \to W$ such that it maps a $w_d \in U$ to the world $w \in W$ such that $d = (w, \Sigma)$. It remains to show that f is surjective and that $a = a^M$.

We need the infinitary conjunctive consistency condition to show that $f : U \to W$ is surjective. To see this pick an arbitrary $w \in W$. We need to find an element $w_d \in U$ which maps under f to w. Hence we have to find a $\Sigma \in \mathsf{Th}(\mathcal{F})$ such that $(w, \Sigma) \in D$. For this define the set $T = \{t \in S \mid w \in b(t)\}$. Because $w \in \bigcap_{t \in T} b(t)$ we can apply the infinitary conjunctive consistency condition to obtain that $\mathsf{cl}\left(\bigcup_{t \in T} a(t)\right)$ is consistence with respect to local consequence on $\mathcal{F}$. This means that $\bigcup_{t \in T} a(t) \not\models \bot$, where $\models$ is the local consequence relation on frames in $\mathcal{F}$. So there is some world v in some Kripke model based on a frame in $\mathcal{F}$ such that $\bigcup_{t \in T} a(t) \subseteq \mathsf{Th}(v)$. Let $\Sigma = \mathsf{Th}(v)$. Because $a(t) \subseteq \Sigma$ for all $t \in T$ it follows from the definition of D that $(w, \Sigma) \in D$.

Lastly, we show that for all $s \in S$ and $\varphi \in \mathcal{M}$

$$\varphi \in a(s) \quad \text{iff} \quad \varphi \in a^M(s).$$

First assume that $\varphi \in a(s)$. To prove $\varphi \in a^M(s)$ we need to show that $v \in I(\varphi)$ for every $v \in b'(s)$. From $v \in b'(s) = f^{-1}[b(s)]$ it follows that $v \in U$ and hence $v = w_d$ for some $d = (w, \Sigma) \in D$. It also follows that $w = f(v) \in b(s)$. By the definition of D we obtain that $a(s) \subseteq \Sigma$ and so $\varphi \in \Sigma$. Because $\Sigma = \mathsf{Th}(w_d)$ it follows that $v = w_d \in I(\varphi)$.

For the other direction take any $\varphi \in a^M(s)$. We want to show that $\varphi \in a(s)$. To do so we apply the infinitary conjunctive covering condition for a suitable covering $\mathcal{T}$ of $b(s)$. Set

$$\mathcal{T} = \{\{t \in S \mid w \in b(t)\} \mid w \in b(s)\}.$$

We first check that this satisfies the antecedent $b(s) \subseteq \bigcup_{T \in \mathcal{T}} \bigcap_{t \in T} b(t)$ of (ICC). This follows immediately from the definition of $\mathcal{T}$ because for any $w \in b(s)$ we have that $w \in \bigcap_{t \in T} b(t)$ for the $T \in \mathcal{T}$ such that $T = \{t \in S \mid w \in b(t)\}$.

By (ICC) it follows that $\bigcap_{T \in \mathcal{T}} \mathsf{cl}\left(\bigcup_{t \in T} a(t)\right) \subseteq a(s)$. So to obtain $\varphi \in a(s)$ it is sufficient to show that for all $w \in b(s)$

$$\varphi \in \mathsf{cl}\left(\bigcup \{a(t) \mid t \in S, w \in b(t)\}\right).$$

To prove this fix an arbitrary $w \in b(s)$ and assume for a contradiction that $\varphi \notin \mathsf{cl}\left(\bigcup \{a(t) \mid t \in S, w \in b(t)\}\right)$. Because cl is defined with respect to local consequence on $\mathcal{F}$ it follows that there is some world v in some Kripke model based on a frame in $\mathcal{F}$ such that $a(t) \subseteq \mathsf{Th}(v)$ for all $t \in S$ with $w \in b(t)$ but $\varphi \notin \mathsf{Th}(v)$. Define $\Sigma = \mathsf{Th}(v)$. Because $a(t) \subseteq \Sigma$ for all $t \in S$ with $w \in b(t)$ it follows by the definition of D that $(w, \Sigma) \in D$. Set $d = (w, \Sigma)$. We then have that $\mathsf{Th}(w_d) = \Sigma = \mathsf{Th}(v)$ and so $\varphi \notin \mathsf{Th}(w_d)$. The latter is equivalent to $w_d \notin I(\varphi)$ which is a contradiction to the assumption that $\varphi \in a^M(s)$ because $w_d \in f^{-1}[b(s)] = b'(s)$. □

The next goal is to use the previous theorem to characterize splitting interpretability with metasemantic models. For this we first define a special kind of Kripke frames that emulates the behavior of the necessity modality in a family of interpretation functions.

7.6.6. Definition. For every set of worlds W define the *two-dimensional frame* $\mathsf{2D}_W$ *over* W to be the modal frame $\mathsf{2D}_W = (W \times W, R)$ over the set of worlds $W^2 = W \times W$ with accessibility relation $R \subseteq W^2 \times W^2$ such that $(w, v)R(w', v')$ iff $v = v'$.

The following proposition states that the behavior of modal sentences relative to the two-dimensional frame over W is indeed the same as relative to a family of interpretation functions that is defined over W:

7.6.7. Proposition. *Let* $(I_w)_{w \in W}$ *be a family of interpretation functions over* W*. Consider the interpretation* $I : \mathsf{At} \to \mathcal{P}(W \times W)$ *over the two-dimensional frame* $\mathsf{2D}_W$ *over* W *that is defined such that for all* $p \in \mathsf{At}$ *and* $v, w \in W$

$$w \in I_v(p) \quad \textit{iff} \quad (v, w) \in I(p).$$

Then this extends such that for all $\varphi \in \mathcal{M}$ *and* $v, w \in W$

$$w \in I_v(\varphi) \quad \textit{iff} \quad (v, w) \in I(\varphi),$$

where on the left side we evaluate using the semantic clause for the necessity modality relative to an interpretation from Section 6.1 and on the right side we evaluate the modality on the Kripke model obtained from adding I *to* $\mathsf{2D}_W$*.*

Proof. This is shown by an induction on the complexity of φ. If φ is an atomic sentence it follows directly from the assumption. The inductive step for the propositional connectives is simple. For the modality one has to compare the semantic clause of the modality on metasemantic models with the semantic clause of the modality on Kripke models. □

The next lemma shows that one can construct a two-dimensional frame from an arbitrary S5-frame. It is similar to Lemma 3.14 in (Fritz 2011).

7.6.8. LEMMA. *Let $F = (W, R)$ be any S5-frame and $A \subseteq W$ non-empty. Then there is a bounded morphism $f : 2\mathsf{D}_W \to F$ such that*

$$A = f[\Delta_W] = \{f(w, w) \in W \mid w \in W\}$$

Proof. We first define for each $a \in A$ a surjective function $s_a : W \to R[\{a\}]$ such that $s_a(a) = a$. This can be done as follows

$$s_a(w) = \begin{cases} w, & \text{if } w \in R[\{a\}], \\ a, & \text{otherwise.} \end{cases}$$

This is well-defined because $a \in R[\{a\}]$ by the reflexivity of R. It is surjective because for every $w \in R[\{a\}]$ it holds that $s_a(w) = w$.

For each $v \in W \setminus A$ we define a surjective function $s_v : W \to R[\{a_0\}]$ with $s_v(v) = a_0$ where $a_0 \in A$ is an arbitrary element of A which exists since A is not empty. The function s_v can be defined as follows

$$s_v(w) = \begin{cases} a_0, & \text{if } w = v \text{ or } w \notin R[\{a_0\}], \\ v, & \text{if } w = a_0 \text{ and } v \in R[\{a_0\}], \\ w, & \text{otherwise.} \end{cases}$$

Again, by the reflexivity of R, we have that $a_0 \in R[\{a_0\}]$ and so this is well-defined. Clearly $s_v(v) = a_0$. The function is surjective because $R[\{a_0\}] \subseteq W$ and a_0 takes the role of v in case $v \in R[\{a_0\}]$.

Now define the bounded morphism $f : 2\mathsf{D}_W \to F, (w, v) \mapsto s_v(w)$. It holds that $A \subseteq f[\Delta_W]$ because $s_a(a) = a$ for all $a \in A$. The other inclusion $A \supseteq f[\Delta_W]$ follows since $s_w(w) \in A$ for all $w \in W$.

It remains to show that f is a bounded morphism. For the forth-condition assume $(w, u)R(w', u')$ in $2\mathsf{D}_W$. By definition this means that $u' = u$ and hence we have to show that $s_u(w)Rs_u(w')$ in F. This follows because the codomain of s_u is either $R[\{u\}]$ or $R[\{a_0\}]$ and the relation R of F is Euclidean. For the back-condition assume that $s_u(w)Rz$ in F. Because $s_u(w) \in R[\{t\}]$, where t is either u or a_0, it follows by the transitivity of R in F that $z \in R[\{t\}]$. Because $s_u : W \to R[\{t\}]$ is surjective there must then be a $w' \in W$ such that $s_u(w') = z$. Hence $f(w', u) = z$ and $(w, u)R(w', u)$ by the definition of $2\mathsf{D}_W$. □

We can now characterize metasemantic splitting interpretability.

7.6.9. THEOREM. *A linguistic behavior $a : S \to \mathcal{PM}$ is metasemantically splitting interpretable with a belief function $b : S \to \mathcal{P}W$ iff it satisfies the conjunctive covering condition* (CC) *and the conjunctive consistency condition* (CCons) *relative to b, where the notion of consequence used in these conditions is local consequence in S5.*

Proof. The proof applies Theorem 7.6.5 to the class $\mathcal{F}$ of all S5-frames, that are frames based on an equivalence relation. Clearly, $\mathcal{F}$ is closed under coproducts and local consequence on $\mathcal{F}$ is local consequence in S5. By Theorem 7.6.5 it follows that a behavior is splitting $\mathcal{F}$-interpretable with some belief function iff it satisfies the infinitary conjunctive covering condition (ICC) and the infinitary conjunctive consistency condition (ICCon) with respect to local consequence in S5. One can show that because local consequence in S5 is compact the infinitary conjunctive covering condition and the infinitary conjunctive consistency condition for local consequence in S5 are equivalent to the conjunctive covering condition (CCons) and the conjunctive consistency condition (CCons) mentioned in the statement of this theorem. Hence for the theorem it remains to be proven that a behavior $a : S \to \mathcal{PM}$ is metasemantically splitting interpretable with a belief function $b : S \to \mathcal{P}W$ iff a is splitting $\mathcal{F}$-interpretable with b.

We first show that metasemantic splitting interpretability implies splitting $\mathcal{F}$-interpretability. So fix some belief function $b : S \to \mathcal{P}W$ and some splitting family of interpretation functions $(W', f', (I_w)_{w \in W'})$ over W. Now consider the linguistic behavior $a^M : S \to \mathcal{PM}$ that is generated by the metasemantic model $M = (W', b', (I_w)_{w \in W'})$ where $b' : S \to \mathcal{P}W'$ is defined such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$. We need to show that a^M is splitting $\mathcal{F}$-interpretable with b. To this aim we define splitting relational model (W'', R, U, g, I) over W such that $a^M = a^{M'}$ for the relational model $M' = (W'', R, U, b'', I)$ where $b'' : S \to \mathcal{P}W''$ is such that $b''(s) = g^{-1}[b(s)]$ for all $s \in S$.

The model M' is based on the two-dimensional frame $\mathsf{2D}_{W'}$ over W'. That is we define the domain $W'' = W' \times W'$ to be the product of W' with itself and the relation $R \subseteq W'' \times W''$ is such that $(v, w)R(v', w'')$ iff $v = v'$. One can easily verify that this R is an equivalence relation and hence $\mathsf{2D}_{W'}$ is in $\mathcal{F}$. The interpretation $I : \mathsf{At} \to \mathcal{P}W''$ is set such that for all $p \in \mathsf{At}$

$$I(p) = \{(v, w) \in W' \times W' \mid w \in I_v(p)\}.$$

This definition satisfies the assumption of Proposition 7.6.7 and hence we obtain that for all formulas $\varphi \in \mathcal{M}$ and worlds $w, v \in W'$

$$w \in I_v(\varphi) \quad \text{iff} \quad (w, v) \in I(\varphi). \tag{7.12}$$

The subset $U \subseteq W''$ is defined as the set $U = \Delta_{W'} = \{(w, w) \in W' \times W' \mid w \in W\}$. The splitting function $g : U \to W$ is such that $g(w, w) = f(w)$. This is surjective because f is surjective.

To prove that $a^M = a^{M'}$ we need to show that for all $s \in S$ and $\varphi \in \mathcal{M}$

$$b'(s) \subseteq D(\varphi) \quad \text{iff} \quad b''(s) \subseteq I(\varphi),$$

where $D(\varphi) = \{w \in W' \mid w \in I_w(\varphi)\}$ is the diagonal proposition of φ relative to the family of interpretations $(I_w)_{w \in W'}$.

First assume that $b'(s) \subseteq D(\varphi)$. We need to show that $b''(s) \subseteq I(\varphi)$. So pick any $(v, w) \in b''(s)$. Because $b''(s) \subseteq U = \Delta_{W'}$ it follows that (v, w) is of the form (w, w) for some $w \in W'$. Because $b''(s) = g^{-1}[b(s)]$ we have that $f(w) = g(w, w) \in b(s)$. So $w \in f^{-1}[b(s)] = b'(s)$. With that assumption that $b'(s) \subseteq D(\varphi)$ we obtain that $w \in I_w(\varphi)$. With (7.12) we obtain that $(v, w) = (w, w) \in I(\varphi)$, which is what we had to show.

For the other direction assume that $b''(s) \subseteq I(\varphi)$. We need to show that $b'(s) \subseteq D(\varphi)$. So take any $w \in b'(s) = f^{-1}[b(s)]$. Then we have that $f(w) \in b(s)$ and hence $g(w, w) \in b(s)$. This shows that $(w, w) \in g^{-1}[b(s)] = b''(s)$. Because $b''(s) \subseteq I(\varphi)$ it follows that $(w, w) \in I(\varphi)$ and so by (7.12) $w \in I_w(\varphi)$. By the definition of $D(\varphi)$ this shows that $w \in D(\varphi)$ as required.

We now prove that splitting $\mathcal{F}$-interpretability implies splitting metasemantic interpretability. So fix some belief function $b : S \to \mathcal{P}W$ and some arbitrary splitting relational model (W', R, U, f, I). This model witnesses the splitting $\mathcal{F}$-interpretability of the linguistic behavior $a^M : S \to \mathcal{P}\mathcal{M}$ that is generated by the relational model $M = (W', R, U, b', I)$ where $b' : S \to \mathcal{P}W'$ is such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$. We have to show that a^M is metasemantically splitting interpretable with b. So we need to find a splitting family of interpretations that interprets a^M with b.

First, let us see that it is sufficient to consider the case where U is not empty. If U is empty then it follows that W is empty because the splitting function $f : U \to W$ is a surjective. It also follows that for that for all $s \in S$ the set $a^M(s) = \mathcal{M}$ is inconsistent and hence we can interpret it with a trivial splitting family of interpretation functions $(U, f, (I)_{w \in U})$ where none of the I_w needs to be defined because U is empty.

Now we can work with the assumption that U is not empty. Consider the S5-frame (W', R) that underlies the interpreting splitting relational model. By Lemma 7.6.8 there is a bounded morphism $g' : W' \times W' \to W'$ from $\mathsf{2D}_{W'}$ to (W', R) which satisfies that $g'[\Delta_{W'}] = U$. We can define an interpretation function $I' : \mathsf{At} \to \mathcal{P}(W' \times W')$ using g' by setting $I'(p) = \{(v, w) \in W' \times W' \mid g'(v, w)\}$. This turns g' into a bounded morphism of Kripke models from the Kripke model obtained from adding the interpretation I' to the frame $\mathsf{2D}_{W'}$ to the Kripke model (W', R, I) that underlies the interpreting splitting relational model. Because bounded morphism of models preserve the truth of formulas it follows that for all $v, w \in W'$ and $\varphi \in \mathcal{M}$

$$(v, w) \in I'(\varphi) \quad \text{iff} \quad g'(v, w) \in I(\varphi). \tag{7.13}$$

We can now define the splitting family of interpretation functions over W that interprets a^M with b. It is defined as $(W', g, (I_w)_{w \in W'})$ where W' is the same domain as in the original splitting relational model (W', R, U, f, I), $g : W' \to W$ is such that $g(w) = f(g'(w, w))$ for all $w \in W'$ and $I_v(p) = \{w \in W' \mid (v, w) \in I'(p)\}$ for all $v \in W'$ and $p \in \mathsf{At}$. It is not obvious that g is well-defined because the domain of f is just U and not all of W'. However the definition works because we know from the construction of g' that $g'[\Delta_{W'}] = U$ and hence $g'(w, w) \in U$ for all $w \in W'$.

From the definition of the family $(I_w)_{w \in W'}$ one can see that it satisfies the assumptions of Proposition 7.6.7. Hence we obtain that for all $w, v \in W'$ and $\varphi \in \mathcal{M}$

$$w \in I_v(\varphi) \quad \text{iff} \quad (v, w) \in I'(\varphi).$$

Connecting to (7.13) yields that

$$w \in I_v(\varphi) \quad \text{iff} \quad g'(v, w) \in I(\varphi). \tag{7.14}$$

To show that $(W', g, (I_w)_{w \in W'})$ interprets a^M with b we need to check that $a^M = a^{M'}$ for the metasemantic model $M' = (W', b'', (I_w)_{w \in W'})$ where $b'' : S \to \mathcal{P}W'$ is defined such that $b''(s) = g^{-1}[b(s)]$ for all $s \in S$. The claim that $a^M = a^{M'}$ is equivalent to the claim that for all $s \in S$ and $\varphi \in \mathcal{M}$

$$f^{-1}[b(s)] \subseteq I(\varphi) \quad \text{iff} \quad g^{-1}[b(s)] \subseteq D(\varphi), \tag{7.15}$$

where $D(\varphi) = \{w \in W' \mid w \in I_w(\varphi)\}$.

For the left-to-right direction assume that $f^{-1}[b(s)] \subseteq I(\varphi)$ and pick any $w \in g^{-1}[b(s)]$. We show that then $w \in D(\varphi)$. From $w \in g^{-1}[b(s)]$ it follows that $g(w) = f(g'(w, w)) \in b(s)$. So $g'(w, w) \in f^{-1}[b(s)]$ and so by assumption $g'(w, w) \in I(\varphi)$. From (7.14) it follows that $w \in I_w(\varphi)$ and hence $w \in D(\varphi)$.

For the right-to-left direction of assume that $g^{-1}[b(s)] \subseteq D(\varphi)$ and pick any $w \in f^{-1}[b(s)]$. We show that then $w \in I(\varphi)$. Because the domain of f is U and $w \in f^{-1}[b(s)]$ it follows that $w \in U$. From the construction of g' we have that $U = g^{-1}[\Delta_{W'}]$ and hence there is some $v \in W'$ such that $g'(v, v) = w$. It follows that $g(v) = f(g'(v, v)) \in b(s)$ because $w \in f^{-1}[b(s)]$. So $v \in g^{-1}[b(s)]$ and so by our assumption $v \in D(\varphi)$. From the definition of $D(\varphi)$ we obtain that $v \in I_v(\varphi)$ and by (7.14) that $g'(v, v) \in I(\varphi)$ which yields $w \in I(\varphi)$ because $w = g'(v, v)$. □

The next goal is to characterize two-dimensional supervaluation interpretability. In the case where W is finite we need that the subject accepts sentence according the modal logic S5.n that is defined as follows:

7.6.10. Definition. For every natural number n define *S5.n* as the smallest normal modal logic that contains S5 and the following axiom λ_n:

$$\lambda_n = \Box \bigvee \{p_1, \ldots, p_n, p_{n+1}\} \to \bigvee \left\{ \Box \bigvee G \mid G \subseteq \{p_1, \ldots, p_n, p_{n+1}\}, |G| \leq n \right\}.$$

So we have for instance that

$$\begin{aligned}
\lambda_1 &= \Box(p_1 \vee p_2) \to (\Box p_1 \vee \Box p_2) \\
\lambda_2 &= \Box(p_1 \vee p_2 \vee p_3) \to (\Box(p_1 \vee p_2) \vee \Box(p_1 \vee p_3) \vee \Box(p_2 \vee p_3)) \\
\lambda_3 &= \ldots
\end{aligned}$$

A different axiomatization of the modal logic S5.n is given by Gärdenfors (1973). I use the axiomatization from Definition 7.6.10 because it makes the similarity to the notion of an n-theory explicit. One can show that a complete S5-theory Σ is a S5.n-theory iff the set of sentences $\Box\Sigma = \{\varphi \in \mathcal{M} \mid \Box\varphi \in \Sigma\}$ is an n-theory in the sense of Definition 7.3.6, with $\mathcal{M}$ instead of $\mathcal{B}$.

In the modal logic S5.n the number of possible worlds is restricted by the finite number n. Every maximally consistent S5.n theory can be witnessed by a model that contains at most n possible worlds. This fact is made precise by the following proposition:

7.6.11. PROPOSITION. *S5.n is sound and strongly complete with respect to the class of all frames whose relation is an equivalence relation in which every equivalence class contains no more than n distinct worlds.*

Proof. This can be proven with the canonical model construction. To check that the axiom λ_n enforce that in the canonical model no world has more than n successors one uses an argument is similar to the proof of Lemma 7.3.7. □

Scroggs (1951) shows that the logics S5.n are the only non-trivial normal modal logics that extend S5. So there is no hope to obtain a counterpart of Proposition 7.6.11 for models in which the number of worlds in an equivalence class are restricted by an infinite cardinality. Such a restriction would however be needed in the characterization of two-dimensional supervaluation interpretability. We solve this problem similarly as in the case of tight interpretability by restricting the language of the subject. The following proposition shows that in this case every maximally consistent S5-theory can be witnessed by a countable model:

7.6.12. PROPOSITION. *If* At *is at most countably infinite then S5 is sound and strongly complete with respect to the class of all frames whose relation is an equivalence relation in which every equivalence class is of at most countable cardinality.*

Proof. The proof goes by applying a selection argument similar to the proof Lemma 7.3.8 to an S5-model that is obtained from the usual strong completeness proof for S5. □

We can now proof the theorem that characterizes two-dimensional splitting interpretability. Its proof is a mixture of the proofs from Theorem 7.5.1 and Lemma 7.6.8.

7.6.13. THEOREM. *Assume that* At *is at most countably infinite. A linguistic behavior* $a : S \to \mathcal{PM}$ *is two-dimensional supervaluation interpretable with a belief function* $b : S \to \mathcal{P}W$ *iff it satisfies the conjunctive covering condition* (CC) *and the conjunctive consistency condition* (CCons) *relative to* b*, where the notion of consequence used in these conditions is local consequence in either S5 if* W *is infinite or in S5.n if there are* n *elements in* W*.*

Proof. The idea of the proof is to apply Theorem 7.6.5 with the class of frames $\mathcal{F}$ that is defined such that it contains all frames (W, R) such that R is an equivalence relation in which all equivalence classes have a cardinality smaller or equal cardinality than W. Clearly $\mathcal{F}$ is closed under coproducts. Hence we obtain by Theorem 7.6.5 that a linguistic behavior is splitting $\mathcal{F}$-interpretable with some belief function iff it satisfies the infinitary conjunctive covering condition (ICC) and the infinitary conjunctive consistency condition (ICCon) with respect to local consequence on $\mathcal{F}$.

By Proposition 7.6.11 we know that if W is finite then S5.n is sound and strongly complete with respect to $\mathcal{F}$. So local consequence on $\mathcal{F}$ is the same as consequence in S5.n. If W is infinite then it follows by Proposition 7.6.12 that S5 is sound and strongly complete with respect to $\mathcal{F}$ because a countable set has smaller or equal cardinality than any infinite set. So in this case local consequence in $\mathcal{F}$ is the same as consequence in S5. We also know that local consequence on $\mathcal{F}$ is compact because it has a strongly complete axiomatization. From this it follows that the infinitary conjunctive covering condition and the infinitary conjunctive consistency condition are equivalent to the conjunctive covering condition (CC) and the conjunctive consistency condition (CCons) mentioned in statement of this theorem. Hence to prove the theorem it remains to be shown that behavior $a : S \to \mathcal{PM}$ is two-dimensional splitting interpretable with a belief function $b : S \to \mathcal{P}W$ iff a is splitting show $\mathcal{F}$-interpretable with b.

To show that two-dimensional supervaluation interpretability implies splitting $\mathcal{F}$-interpretability consider the behavior a^M generated by some two-dimensional supervaluation model $M = (W, b, \mathcal{I})$. We have to show that a^M is splitting $\mathcal{F}$ interpretable with b. So we need to find a splitting relational model (W', R, U, f, I') such that (W', R) is in $\mathcal{F}$ and $a^M = a^{M'}$ for the relational model $M' = (W', R, b', I')$, where $b' : S \to \mathcal{P}W'$ is defined such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$.

The domain of M' is the set $W' = \mathcal{I} \times W \times W$. The relation $R \subseteq W' \times W'$ is defined such that $(I, v, w)R(\tilde{I}, v', w')$ iff $I = \tilde{I}$ and $v = v'$. Clearly this is an equivalence relation and every equivalence class has cardinality W. Hence (W', R) is in $\mathcal{F}$. The set $U \subseteq W'$ is defined as $U = \{(I, v, w) \in W' \mid v = w\}$. The splitting function $f : U \to W$ maps an element (I, w, w) to $f(I, w, w) = w$. Clearly this is surjective. Lastly, we define the interpretation $I' : \mathsf{At} \to \mathcal{P}W'$ such that $I'(p) = \{(I, v, w) \in W' \mid (v, w) \in I'(p)\}$.

We need the fact that our definition of the interpretation I' is such that $v, w \in W$, $I \in \mathcal{I}$ and $\varphi \in \mathcal{M}$

$$(v, w) \in I(\varphi) \quad \text{iff} \quad (I, v, w) \in I'(\varphi). \tag{7.16}$$

This is proven by an induction on the complexity of φ. The base case follows immediately from the definition of I' and the inductive cases for the propositional connectives are standard. The case for the modality follows from comparing the semantic clause of the modality relative to a two-dimensional interpretation with the definition of the relation R'.

We next show the required $a^M = a^{M'}$. By considering the definition of the behaviors generated by a two-dimensional supervaluation model and by a relational model one finds that this amounts to the claim that for all $s \in S$ and $\varphi \in \mathcal{M}$

$$b(s) \subseteq (I(\varphi))^d \text{ for all } I \in \mathcal{I} \quad \text{iff} \quad f^{-1}[b(s)] \subseteq I'(\varphi).$$

For the left-to-right direction assume that $b(s) \subseteq (I(\varphi))^d$ for all $I \in \mathcal{I}$ and take any $(I, w, w) \in f^{-1}[b(s)]$. We want to show that $(I, w, w) \in I'(\varphi)$. From $(I, w, w) \in f^{-1}[b(s)]$ it follows that $w \in b(s)$. By the assumption that $b(s) \subseteq (I(\varphi))^d$ it follows that $(w, w) \in I(\varphi)$ which by (7.16) implies that $(I, w, w) \in I'(\varphi)$.

For the right-to-left direction assume that $f^{-1}[b(s)] \subseteq I'(\varphi)$. We want to show that $b(s) \subseteq (I(\varphi))^d$ for all $I \in \mathcal{I}$. So fix any $I \in \mathcal{I}$ and consider some $w \in b(s)$. We need that $w \in (I(\varphi))^d$. Because $w \in b(s)$ it follows that $(I, w, w) \in f^{-1}[b(s)]$. Hence $(w, w, I) \in I'(\varphi)$. By (7.10) it follows that $(w, w) \in I(\varphi)$ which yields $w \in (I(\varphi))^d$.

We now show that splitting $\mathcal{F}$-interpretability implies two-dimensional supervaluation interpretability. That a behavior a is splitting $\mathcal{F}$-interpretable with a belief function $b : S \to \mathcal{P}W$ means that there is some splitting relational model (W', R, U, f, I) such that (W', R) is in $\mathcal{F}$ and $a = a^M$ for the relational model (W', R, b', I), where $b' : S \to \mathcal{P}W'$ is such that $b'(s) = f^{-1}[b(s)]$ for all $s \in S$. To show that a is two-dimensional supervaluation interpretable we have to find a two-dimensional supervaluation $\mathcal{I}$ such that $a^M = a^{M'}$ for the two-dimensional supervaluation model $M' = (W, b, \mathcal{I})$.

To define I first take an index set J that has the cardinality of the supremum of the cardinalities of all the sets $f^{-1}[\{w\}]$ for some $w \in W$. That $f : U \to W$ is surjective implies that $f^{-1}[\{w\}]$ is not empty for every $w \in W$. Hence we can choose for every $w \in W$ a surjective function $g_w : J \to f^{-1}[\{w\}]$.

We can assume without loss of generality that the set J is not empty. If J was empty then J it would follows by the surjectivity of f that W is empty. This implies that $a^M(s)$ is inconsistent for every $s \in S$. But then a^M is interpretable by the two-dimensional supervaluation model $(W, b, \mathcal{I})$ where $\mathcal{I} = \{I\}$ and $I : \mathsf{At} \to \mathcal{P}(W \times W)$ is such that $I(p) = \emptyset = W \times W$ for all $p \in \mathsf{At}$.

Next consider any $w' \in f^{-1}[\{w\}]$ for some fixed $w \in W$. From the choice of $\mathcal{F}$ we know that for any such w' the set $R[\{w'\}]$ has a cardinality that is smaller than W. Moreover $R[\{w'\}]$ is not empty because R is reflexive. So there exists a surjective function $h_{w'} : W \to R[\{w'\}]$. We can also assume that $h_{w'}(w) = w'$ because if this was not the case we could make it the case by swapping the value of w under $h_{w'}$ with the value of some element in $h_{w'}^{-1}[\{w'\}]$, which exists because $h_{w'}$ is surjective.

We can now define the two-dimensional supervaluation $\mathcal{I} = \{I_j \mid j \in J\}$, where for any $j \in J$ the two-dimensional interpretation $I_j : \mathsf{At} \to \mathcal{P}(W \times W)$ is such that for all $p \in \mathsf{At}$

$$I_j(p) = \{(v, w) \mid h_{g_v(j)}(w) \in I(p)\}.$$

The set $\mathcal{I}$ is not empty because J is assumed to be non-empty.

From the definition of $\mathcal{I}$ it follows for all $j \in J$, $v, w \in W$ and $\varphi \in \mathcal{M}$

$$(v, w) \in I_j(\varphi) \quad \text{iff} \quad h_{g_v(j)}(w) \in I(\varphi). \tag{7.17}$$

This is proven by an induction on the complexity of φ. In the base case φ is an atomic sentence and hence the claim follows from the definition of the interpretation I_j for some $j \in J$. The inductive step is obvious in the cases where the main connective of φ is propositional. We only give the case where φ is a modal formula of the form $\Box\psi$.

First consider the left-to-right direction. So assume that $(v, w) \in I_j(\Box\psi)$. We need to show that $h_{g_v(j)}(w) \in I(\Box\psi)$. The relevant semantic clause here is the one for the Kripke model (W, R, I). So we have to show the for every $u \in W$ such that $h_{g_v(j)}(w)Ru$ it is the case that $u \in I(\psi)$. Fix any such u. We now first show that $u \in R[\{g_v(j)\}]$. This follows from the transitivity of R and the fact that $g_v(j)Rh_{g_v(j)}(w)$, where the latter is the case because $h_{g_v(j)} : W \to R[\{g_v(j)\}]$ is defined to map onto $R[\{g_v(j)\}]$. From $u \in R[\{g_v(j)\}$ it follows by the subjectivity of $h_{g_v(j)}$ that there is some $z \in W$ such that $h_{g_v(j)}(z) = u$. So it remains to show that $h_{g_v(j)}(z) \in I(\psi)$ which by the induction hypothesis is equivalent to $(v, z) \in I_j(\psi)$. But this follows from the assumption that $(v, w) \in I_j(\Box\psi)$ and the semantic clause of the necessity modality on two-dimensional interpretations.

For the right-to-left direction of (7.17) with $\varphi = \Box\psi$ assume that $h_{g_v(j)}(w) \in I(\Box\psi)$. We want to show that $(v, w) \in I_j(\Box\psi)$. By the semantic clause of the modality relative to two-dimensional interpretations this means that we need to show for any $u \in W$ that $(v, u) \in I_j(\psi)$ which by the induction hypothesis amounts to $h_{g_v(j)}(u) \in I(\psi)$. So fix an arbitrary $u \in W$. Because $h_{g_v(j)}$ is defined to map into the set $R[\{g_v(j)\}]$ we have that $h_{g_v(j)}(u) \in R[\{g_v(j)\}]$ and that $h_{g_v(j)}(w) \in R[\{g_v(j)\}]$. Because R is Euclidean it follows that $h_{g_v(j)}(w)Rh_{g_v(j)}(u)$. Now we get the needed $h_{g_v(j)}(u) \in I(\psi)$ from the assumption that $h_{g_v(j)}(w) \in I(\Box\psi)$ using the semantic clause for the modality on the Kripke model (W, R, I).

It remains to show that with the above definition of $\mathcal{I}$ we have that $a^M = a^{M'}$. By unfolding the respective definitions of the behavior generated by a relational model and by a two-dimensional supervaluation model on can see that this amounts to showing that for all $s \in S$ and $\varphi \in \mathcal{M}$

$$f^{-1}[b(s)] \subseteq I(\varphi) \quad \text{iff} \quad b(s) \subseteq (I_j(\varphi))^d \text{ for all } j \in J. \tag{7.18}$$

For the left-to-right direction of (7.18) assume that $f^{-1}[b(s)] \subseteq I(\varphi)$. We then take any $w \in b(s)$ for which we want to show that $w \in (I_j(\varphi))^d$ for all $j \in J$. So fix an arbitrary $j \in J$. We need to show that $w \in (I_j(\varphi))^d$ which by the definition of the diagonal amounts to showing $(w, w) \in I_j(\varphi)$. From the definition of $g_w : J \to f^{-1}[\{w\}]$ it follows that $g_w(j) \in f^{-1}[\{w\}]$. Hence we can consider the function $h_{g_w(j)}$ as defined above for which it holds that $h_{g_w(j)}(w) = g_w(j)$. Because $g_j(j) \in f^{-1}[\{w\}]$ and $w \in b(s)$ it follows that $h_{g_w(j)}(w) \in f^{-1}[b(s)]$.

From applying the assumption $f^{-1}[b(s)] \subseteq I(\varphi)$ we get that $h_{g_w(j)}(w) \in I(\varphi)$. With (7.17) we obtain the required $(w, w) \in I_j(\varphi)$.

For the right-to-left direction of (7.18) assume that $b(s) \subseteq (I_j(\varphi))^d$ for all $j \in J$ and consider any $w' \in f^{-1}[b(s)]$. We need to show that $w' \in I(\varphi)$. First define $w \in f(w')$. Since $g_w : J \to f^{-1}[\{w\}]$ is surjective and $w' \in f^{-1}[\{w\}]$ there is some $j \in J$ such that $g_w(j) = w'$. Because $w' \in f^{-1}[b(s)]$ it follows that $w = f(w') \in b(s)$ which by assumption entails that $w \in (I_j(\varphi))^d$. By definition this is equivalent to $(w, w) \in I_j(\varphi)$. Using (7.17) we obtain that $h_{g_w(j)}(w) \in I(\varphi)$. It follows that $w' \in I(\varphi)$ because we can show that $h_{g_w(j)}(w) = w'$. This holds because $g_w(j) = w'$ and $h_{w'}(w) = w'$ by the special property of the function $h_{w'} : W \to R[\{w'\}]$. □

I end this section with an example which demonstrates the difficulties for finding conditions that characterize the notion of interpretability suggested at the end of Section 6.2. I start by formally defining the notion of interpretability that is suggested there.

7.6.14. Definition. A *relational supervaluation model* $(W, R, b, \mathcal{I})$ consists of a supervaluation model $(W, b, \mathcal{I})$ together with an equivalence relation $R \subseteq W \times W$ on W.

The *linguistic behavior* $a^M : S \to \mathcal{PM}$ *generated* by a relational supervaluation model $M = (W, R, U, b, I)$ is defined such that for all situations $s \in S$

$$a^M(s) = \{\varphi \in \mathcal{M} \mid b(s) \subseteq I(\varphi) \text{ for all interpretations } I \in \mathcal{I}\},$$

where an interpretation $I : \mathsf{At} \to \mathcal{P}W$ extends to $\mathcal{M}$ with the usual semantic clause on Kripke frames, that is,

$$I(\Box\varphi) = \{w \in W \mid v \in I(\varphi) \text{ for all } v \in W \text{ with } wRv\}.$$

A supervaluation $\mathcal{I}$ over W and an equivalence relation $R \subseteq W \times W$ *interpret a linguistic behavior* $a : S \to \mathcal{PM}$ *with a belief function* $b : S \to \mathcal{P}W$ if $a = a^M$ for the relational supervaluation model $M = (W, R, b, \mathcal{I})$.

A linguistic behavior $a : S \to \mathcal{PM}$ is *relationally supervaluation interpretable with a belief function* $b : S \to \mathcal{P}W$ if there exists some supervaluation $\mathcal{I}$ over W and an equivalence relation $R \subseteq W \times W$ on W that interpret a with b.

The following example gives an idea of the difficulties in finding necessary and sufficient conditions for relational supervaluation interpretability:

7.6.15. Example. Take a domain $W = \{w, v\}$ with two worlds and assume that $\mathsf{At} = \{p, q\}$. Also consider two situations $S = \{s, t\}$ and a belief function $b : S \to \mathcal{P}W$ such that $b(s) = \{w\}$ and $b(t) = \{v\}$.

We define a two-dimensional interpretation function $I : \mathsf{At} \to \mathcal{P}W$ such that $I(p) = \{(w, w), (w, v), (v, w)\}$ and $I(q) = \{(w, w), (v, w)\}$. This two-dimensional interpretation is also specified by the table

	w	v
w	$p, \ q$	$p, \neg q$
v	$p, \ q$	$\neg p, \neg q$

We can turn this into the singleton two-dimensional supervaluation $\mathcal{I} = \{I\}$.

Consider the linguistic behavior $a = a^M$ generated by the two-dimensional supervaluation model $(W, b, \mathcal{I})$. By definition this behavior is two-dimensional supervaluation interpretable with b. It is also metasemantically splitting interpretable with b because two-dimensional supervaluation interpretability implies metasemantic splitting interpretability. One can also see this directly by considering the identity splitting function and the family of interpretations that is encoded in the table above.

The behavior a is however not relationally supervaluation interpretable. To see this first check with the table above that $\{\Box p, q, \neg\Box q\} \subseteq a^M(s)$ and $\neg p \in a^M(t)$. Now assume for a contradiction that there is a supervaluation $\mathcal{I}$ over W and an equivalence relation $R \subseteq W \times W$ on W that interpret a. Consider an arbitrary interpretation $I \in \mathcal{I}$. Because $\neg\Box q \in a(s)$ it follows that $w \in I(\neg\Box q)$. Hence there is some world u with wRu such that $u \notin I(q)$. The world u can not be identical to w because $q \in a(s)$ and hence $w \in I(q)$. So u must be identical to v. But this is not possible because $\Box p \in a(s)$, hence $w \in I(\Box p)$ and so $u \in I(p)$, which contradicts the $v \notin I(p)$ that follows from $\neg p \in a(t)$.

More intuitively, the reason why a is not relationally supervaluation interpretable is that we would need an additional counterfactual world in the equivalence class of u that can not be identical to v. But there can not be such a world because we have only two worlds in the domain.

To characterize relational supervaluation interpretability we need to exclude behaviors such as a. For this we have to find conditions that impose an additional counterfactual cardinality constraint, which is stronger than the constraint for two-dimensional supervaluation interpretability. The size of W does not just restrict the number of counterfactual worlds that exist from the perspective of a given doxastic alternative but it is an upper bound on the total number of counterfactual worlds that exist from the perspectives of all the doxastic alternatives together.

Chapter 8

Conclusions

This last chapter contains concluding remarks that concern the whole thesis. In Section 8.1 I express some thoughts about how one might think about the kind of representation results given in this thesis. Section 8.2 addresses again the choice between disquotational and metasemantic acceptance which is the common theme of the later chapters in this thesis. I distinguish between three different notions of meaning in the formal framework and suggest that the choice between disquotational and metasemantic acceptance is a choice about how these notions of meaning relate to each other. In Section 8.3 I sketch various ideas for further work that extends the setting of this thesis.

8.1 What is this good for?

The main contribution of this thesis is to develop a formal account of radical interpretation. To this aim I introduce a notion of linguistic behavior, I define when a linguistic behavior is interpreted by some formal model for belief and meaning and prove a representation result that characterizes the class of interpretable behaviors.

The work in this thesis improves upon existing accounts of radical interpretation in that it formalizes the problem of interpretation and its solution to the problem within the possible world framework. This shows that the problem of radical interpretation can be treated on a level of mathematical sophistication that is comparable to that used in decision theory.

Because they are formulated in the possible world framework these results on radical interpretation are also relevant for the understanding of possible world models. This thesis does not introduce any new structures for representing beliefs or meanings. It concerns formal representations of beliefs and meanings that have all been around since at least the seventies. Still, I think that the results from this thesis are valuable for doxastic logic and formal semantics.

The representation results of this thesis provide a similar foundation for doxastic logic as the representation results in decision theory provide for probability theory. By also considering the meaning of sentences in the language of the subject one can reduce the beliefs of the subject, as they are modeled in doxastic logic, to her linguistic behavior. This provides a theoretical foundation for the models that are used in doxastic logic. However, I do not expect that the results of this thesis are actually going to be useful for the kind of work that doxastic logicians usually do. In most applications of doxastic logic it is perfectly fine to assume that we know what propositions the sentences in the language of the subject express. Hence one can work in a setting, like the one of Section 2.2, where the interpretation function is fixed in advance.

The setting of this thesis might also influence how one thinks about formal semantics. Commonly, formal semantics is understood as being about the recursive assignment of meanings to sentences in natural language. This view is expressed for instance, with a narrow conception of meaning as truth conditions, by Heim and Kratzer (1998, ch. 1) or, with a less restrictive conception of meaning, by Yalcin (2014). A difficulty with this view is that meaning is itself a theoretical concept of formal semantics. For this reason it might be difficult to judge when a semantic theory succeeds at assigning meanings to sentences because there is no theory-independent standard for what meanings are. This point is developed extensively by Stokhof (2014, sec. 4) while discussing the different conceptions of meaning in dynamic semantics.

The representation results of this thesis suggest a broader conception of formal semantics that also incorporates parts of what is sometimes called formal pragmatics: Formal semantics is about linguistic behavior. As a consequence semantic theories should be evaluated on how well they represent actual linguistic behavior. Such a broader conception of semantics has the advantage that it grounds meaning in the less theoretical and more empirical notion of linguistic behavior. The contribution of this thesis is to formalizes the connection between a compositional semantic theory of meaning and linguistic behavior in the most basic and simple case of the classical semantics for the propositional connectives and the necessity modality.

8.2 Three senses of "meaning"

In the latter three chapters of this thesis, I investigate the difference between the disquotational and metasemantic acceptance principles and compare the different accounts of interpretation that they give rise to. In this section I take a more conceptual view on the difference between the principles. I distinguish between three senses in which one might use "meaning", and related words, when describing the models used in this thesis. This distinction contributes to the understanding of the difference between disquotational and metasemantic acceptance.

The first sense of "meaning" I call metaphysical meaning. The *metaphysical meaning* of a sentence at a world is the structure that is associate to the world in the model to represent the semantic facts that determine the meaning of that sentence. I call this metaphysical meaning because if we think of the formal models as representing reality as it is then metaphysical meaning captures what the semantic facts about are about.

The second sense of "meaning" I call interpretational meaning. The *interpretational meaning* of a sentence in a situation is the part of the model , other than the belief set of the subject, that is needed in the definition of the behavior generated to determine whether the subject accepts the sentence in the situation. Or, to turn this around: Interpretational meaning is the thing that we need to interpret the subject's linguistic behavior in some situation.

The third sense of "meaning" is compositional meaning. The *compositional meaning* of a sentence is the structure that the recursive semantic clauses operate on. It depends on the type of models that we are considering what kind of objects, such as for instance worlds, situations or interpretations, the notion of compositional meaning needs to be relativized to.

Similar distinctions between different notions of meaning already appear in the literature. The distinction between metaphysical and interpretational meaning mirrors the spirit Lewis' (1975) thesis and antithesis about what a language is. The distinction between interpretational meaning and compositional meaning is similar to a distinction that has been made starting from Lewis (1980) between the objects of the propositional attitudes and semantic values (see Yalcin 2014, sec. 2.2 for references). The definition of the behavior generated by some model compares interpretational meaning to the belief set of the subject, hence interpretational meaning is similar to the objects of belief. Semantic values are usually taken to be the objects on which the recursive semantic clauses are defined and hence correspond to compositional meaning.

I now explain how these loose characterization of the different notions of meaning apply to two-dimensional supervaluation models, assuming disquotational acceptance, and to metasemantic models, assuming metasemantic acceptance. First consider a two-dimensional supervaluation model $(W, b, \mathcal{I})$.

The two-dimensional supervaluation $\mathcal{I}$ of the model two-dimensional supervaluation model $(W, b, \mathcal{I})$ represents the semantic facts that obtain at the actual world. With the above characterization of metaphysical meaning it follows the metaphysical meaning of a sentence at the actual world is given by the two-dimensional supervaluation $\mathcal{I}$ of the model. One might think of the metaphysical meaning of a sentence φ at the actual world as the set $\{I(\varphi) \mid I \in \mathcal{I}\}$ of all two-dimensional meanings that the sentence expresses according to one of the two-dimensional interpretation functions in the supervaluation.

To see what the interpretational meaning of a sentence at some world in some two-dimensional supervaluation model is we need to check in Definition 6.4.5 of the behavior generated by a two-dimensional supervaluation model what part of the model is used to determine whether the subject accepts some sentence in some situation. It seems that every interpretation function in the two-dimensional supervaluation of the model might influence whether the subject accepts some sentence. Hence in two-dimensional supervaluation models the interpretational meaning of some sentence in some situation is the same as its metaphysical meaning at the world of the situation.

One might also apply the above characterization of interpretational meaning more strictly and observe that whether the subject accepts some sentence in some situation depends just on the diagonal propositions $\{(I(\varphi))^d \subseteq b(s) \mid I \in \mathcal{I}\}$ that the sentence expresses according to the two-dimensional supervaluations. In this case the metaphysical meaning of the sentence is just the set of diagonal propositions that it expresses according to the interpretation functions in the supervaluation. With this stricter reading it would no longer hold that interpretational meaning of some sentences is the same as its metaphysical meaning. But it would still be the case that the interpretational meaning of some sentence is determined by its metaphysical meaning. For the discussion below it does not matter which of the two reading of the characterization of interpretational meaning we choose. Hence I continue by assuming the simpler reading on which interpretational meaning is the same as metaphysical meaning.

To determine the notion of compositional meaning in two-dimensional supervaluation models first note that the semantic clauses for the propositional connectives and the necessity modality are defined relative to a two-dimensional interpretation function and not relative to a supervaluation. Hence, the compositional meaning of some sentence is relative to the two-dimensional interpretation in the two-dimensional supervaluation that we are considering. Moreover, the semantic clauses never shift the first component of a pair of worlds relative to which some modal formula is evaluated. Therefore we can relativize compositional meaning further to a possible world which we think of as the fixed first component. The compositional meaning of some sentence φ relative to a two-dimensional interpretation I in the supervaluation and relative to some world v is then the proposition $\{w \in W \mid (v, w) \in I(\varphi)\}$.

Let me give an example that illustrates how these notions apply to two-dimensional supervaluation models. I again use the example from Chapter 5 but adapt it to two-dimensional supervaluations. The domain of the model is the set $W = \{w_{1.6}, w_{1.8}, w_{2.0}\}$ such that at w_s the man that the subject is talking about is s meters tall. We consider the sentence "The man is tall." for which we use the letter p. The meaning of p in English is presumably captured by the two-dimensional supervaluation $\mathcal{I} = \{I, I'\}$, where the two-dimensional interpretation functions I is encoded in following table on the left and I' is encoded in the following table on the right:

	$w_{1.6}$	$w_{1.8}$	$w_{2.0}$
$w_{1.6}$	$\neg p$	p	p
$w_{1.8}$	$\neg p$	p	p
$w_{2.0}$	$\neg p$	p	p

	$w_{1.6}$	$w_{1.8}$	$w_{2.0}$
$w_{1.6}$	$\neg p$	$\neg p$	p
$w_{1.8}$	$\neg p$	$\neg p$	p
$w_{2.0}$	$\neg p$	$\neg p$	p

The two-dimensional supervaluation $\mathcal{I}$ is the adaption of the supervaluation from the example in Section 5.2, assuming that the meaning of the sentence p extends in the most natural way to two-dimensional interpretations.

The metaphysical meaning of the sentence “The man is tall.” is given by its values in the supervaluation $\mathcal{I}$ that are encoded in the two tables above.

The interpretational meaning of “The man is tall.” is the same as the metaphysical meaning, and hence is also given by the two tables above. With the more restrictive reading of the characterization of interpretational meaning that I mention above we would have that the interpretational meaning of “The man is tall.” is determined by the entries on diagonals of the two-tables above. This corresponds to the set of diagonal propositions $\{\{w_{1.8}, w_{2.0}\}, \{w_{2.0}\}\}$ and is precisely the information that is captured by the supervaluation from the example in Section 5.2.

The compositional meaning of “The man is tall.” is relative to a two-dimensional interpretation and to a possible world. In the representation of the supervaluation $\mathcal{I}$ in the two tables above this amounts to fixing one of the tables and then a row inside of the table. The intuitive reason for this is that the semantic clause for the necessity modality operates only along a row of one of the tables. For instance we have that the metaphysical meaning of “The man is tall.” relative to the two-dimensional interpretation I and the world $w_{1.8}$ is the proposition $\{w_{1.8}, w_{2.0}\}$. In the example the compositional meaning of this sentence does not change relative to the world that we are considering because in both tables above all rows are identical. For an example of a sentence that has a compositional meaning that depends on the possible worlds the reader can consider the sentence “Water is H_2O.” relative to the two-dimensional interpretation that is encoded in the table on page 102 in Section 6.4

Next, I explain how the above characterizations of the three notions of meaning apply to a metasemantic model $(W, b, (I_w)_{w \in W})$, assuming metasemantic acceptance.

The metaphysical meaning of a sentence φ at a world w is the proposition $I_w(\varphi)$. This follows from the explanation in Section 4.1 of how semantic facts are represented in a metasemantic model.

To see what interpretational meaning of a sentence in some situation is we have to consider Definition 4.3.1 of the behavior that is metasemantically generated by a metasemantic model. Whether the subject accepts some sentence in some situation depends on the interpretation functions that are associated to the doxastic alternatives of the subject in that situation. Hence the interpretational meaning of a sentence φ in a situation s is the set $\{I_w(\varphi) \subseteq W \mid w \in b(s)\}$ of all the proposition that φ expresses according to some world in the belief set of the subject in the situation s. Less formally, one can say that the interpretational meaning of some sentence in some situation is determined by what the subject believes in the situation about the proposition that the sentence expresses according to the semantic facts. Hence the interpretational meaning of some sentence is determined by the beliefs of the subject about the metaphysical meaning of that sentence.

The compositional meaning of some sentence is the structure that the semantic clauses operate on when the sentence occurs embedded in a more complex sentences. Hence consider how we employ the semantic clauses for the propositional connectives, from Section 1.2 and the necessity modality, from Section 6.1, in the case of metasemantic models. They operate on the proposition that some sentence expresses according to some interpretation function, which in metasemantic models is always relative to a possible world. So compositional meaning needs to be relativized to possible worlds. The compositional meaning of a sentence φ at some world w is the proposition $I_w(\varphi)$ that the sentence expresses according to the semantic facts at that world. Hence, in metasemantic models the compositional meaning of a sentence is the same as its metaphysical meaning.

Let me explain in an example how these notions apply to metasemantic models. I use a metasemantic model that is similar to the one discussed in Section 5.1. Take the domain $W = \{w_{1.6}, w'_{1.6}, w_{1.8}, w'_{1.8}, w_{2.0}, w'_{2.0}\}$ such that at w_s and w'_s the man is s meters high and at the w_s worlds the sentence "The man is tall." is true of men that are at least 1.8 meters tall whereas at the w'_s worlds it applies only to men that are taller than 1.8 meters. We use the letter p for the sentence "The man is tall.". This semantic facts about p can be represented by the family of interpretation functions given in the following table:

	$w_{1.6}$	$w'_{1.6}$	$w_{1.8}$	$w'_{1.8}$	$w_{2.0}$	$w'_{2.0}$
$w_{1.6}$	$\neg p$	$\neg p$	p	p	p	p
$w'_{1.6}$	$\neg p$	$\neg p$	$\neg p$	$\neg p$	p	p
$w_{1.8}$	$\neg p$	$\neg p$	p	p	p	p
$w'_{1.8}$	$\neg p$	$\neg p$	$\neg p$	$\neg p$	p	p
$w_{2.0}$	$\neg p$	$\neg p$	p	p	p	p
$w'_{2.0}$	$\neg p$	$\neg p$	$\neg p$	$\neg p$	p	p

This family of interpretation functions is an extension of the one used in the example from Section 5.1, where I have added the worlds $w'_{1.6}$ and $w'_{2.0}$. To obtain a metasemantic model we also need a belief function. Let us use the function $b : S \to \mathcal{P}W$ with $S = \{s\}$ and $b(s) = \{w_{1.8}, w'_{1.8}\}$. Hence, there is only one situation s, in which the subject believes that the man is 1.8 meters tall but she is uncertain whether p applies to men of that height.

The metaphysical meaning of "The man is tall." depends on the possible world in the model. At the worlds w_s it is the proposition $\{w_{1.8}, w'_{1.8}, w_{2.0}, w'_{2.0}\}$

that the man is at least 1.8 meters tall whereas at the worlds w'_s it is the proposition $\{w_{2.0}, w'_{2.0}\}$ that the man is taller than 1.8 meters. In the table above the metaphysical meaning of the sentence at some world corresponds to the row indexed by this world. The compositional meaning of a sentence at a world is the same as its metaphysical meaning at the world. Hence the compositional meaning of a sentence at a world also corresponds to the row in the table that is indexed by the world.

The interpretational meaning of “The man is tall.” in some situation is given by what the subject believes in the situation about the metaphysical meaning of “The man is tall.”. For the situation s, where the subject has the belief set $\{w_{2.0}, w'_{2.0}\}$, this corresponds to the rows indexed by $w_{1.8}$ and $w'_{1.8}$ in the table above.

I continue by discussing the theoretical function of the different notions of meaning and explain how they relate to the distinction between disquotational and metasemantic acceptance.

The notions of interpretational meaning and of compositional meaning are both constrained by interpretation. Interpretational meaning relates to the subject's acceptance of sentences and compositional meaning relates to how sentences behave if they occur as parts of more complex sentences that the subject accepts.

The discussion from Chapters 5 and 6 suggests that the constraints on interpretational and compositional meaning are such that they can not be identical. To account for the problem of vagueness we need that interpretational meaning can leave it undetermined whether a sentence is true or false in a situation where the subject has complete information about the basic facts. But we need to keep a two-valued notion of compositional meaning if we do not want to loose classical logic. To account for necessity a priori we use a two-dimensional approach that disassociates the counterfactual dependence of truth-values that is captured by the compositional meaning and determines the behavior of sentences in the scope of the necessity modality from the epistemic dependence of truth-values that is captured with interpretational meaning and determines what sentences the subject accepts in what situations. As a consequence we have that in both two-dimensional supervaluation models and metasemantic models the interpretational meaning of a sentence is distinct from its compositional meaning.

The formal relation between interpretational and compositional meaning is similar in two-dimensional supervaluation models and in metasemantic models. In both cases the interpretational meaning of some sentence is composed of multiple compositional meanings of the sentence and in both cases compositional meaning is associated to metaphysical necessity that is distinct from the epistemic necessity that is relevant for interpretational meaning. In metasemantic models this relation between interpretational and compositional meaning is implemented by the doxastic structure that relates the subject's beliefs about the semantic facts, which determine interpretational meaning, to the semantic facts, which determine compositional meaning. In two-dimensional supervaluation models this relation between interpretational and compositional meaning is explicitly implemented with a more complex notion of semantic facts, that is rich enough to determine interpretational meaning, and to contain multiple compositional meanings.

Metaphysical meaning is not as directly constrained by interpretation as interpretational and compositional meaning. It is a more theoretical notion. What metaphysical meaning is depends on a decision of the modeler about what kind of semantic structure we want to associate with the possible worlds in our models. This decision is made differently for two-dimensional supervaluation models than for metasemantic models. In two-dimensional supervaluation models the metaphysical meaning of a sentence is identified with or at least determines the interpretational meaning of the sentence. In metasemantic models, assuming the metasemantic acceptance principle, the metaphysical meaning of a sentence is identified with its compositional meaning.

Because two-dimensional supervaluation models are developed to be used with the disquotational acceptance principle and metasemantic models are best used with the metasemantic acceptance principles this gives us a different perspective on the distinction between disquotational and metasemantic acceptance. Simplifying a little we might say that disquotational acceptance identifies the metaphysical meaning with interpretational meaning whereas metasemantic acceptance identifies metaphysical meaning with compositional meaning.

The distinction between metaphysical interpretational and compositional meaning can be extended to other related notions such as the notion of a language and the notion of semantics. I continue by discussing these two cases.

Corresponding to the different sense of "meaning" there are also different senses of "language". A metaphysical language associates sentences with their metaphysical meaning, an interpretational language associates sentences with their interpretational meanings and a compositional language associates sentences with their compositional meaning.

With both acceptance principles we have that there is a distinction between the notions of an interpretational language and the notion of a compositional language. A compositional language is an assignment of sentences to some relatively simple notion of meaning on which the recursive semantic clauses operate. It is the kind of thing that is studied in formal semantics. An interpretational language is a more complex object that is constructed by putting different compositional languages together. It is the kind of thing that we use to interpret the subject. It is also most plausible to think that an interpretational language is denoted by an expression such as "the language of the subject", because in cases of semantic indeterminacy or uncertainty there is no single compositional language of the subject.

A similar point applies to the formal counterparts of languages in the sense in which English or Dutch are languages. A language in this sense most plausibly corresponds to the interpretational language of all of its speakers. With disquotational acceptance we would have that for instance English is given by a two-dimensional supervaluation that can be used to interpret speakers of English. For this it is necessary to assume that one single supervaluation can be used to interpret all speakers of English. With the metasemantic acceptance principle English would be given by the beliefs about semantic facts that are shared among speakers of English. One needs to assume that these beliefs are sufficiently homogeneous.

Analogously to the notion of metaphysical meaning we have that the notion of a metaphysical language depends on the acceptance principle that we are using. With disquotational acceptance a metaphysical language is the same as an interpretational language. With metasemantic acceptance it is the same as a compositional language.

The distinction between an interpretational sense and a compositional sense also applies to the notion of semantics. Interpretative semantics is the study of interpretational meaning and compositional semantics is the study of compositional meaning. It corresponds to the distinction between a narrower and a wider sense of formal semantics at the end of the previous section. Compositional semantics is formal semantics in the narrow sense that concerns the recursive assignment of meanings to sentences. Interpretative semantics is formal semantics in the wider sense that concerns the interpretation of linguistic behavior.

It is not clear whether there is an interesting conception of metaphysical semantics, because metaphysical meaning is a theory internal notion. But the concept of metaphysical semantics can explain the sense in which metasemantic models and the metasemantic acceptance principle are metasemantic. Metasemantic models are metasemantic because they represent the different metaphysical semantics of different worlds and metasemantic acceptance is metasemantic because it depends on the subject's beliefs about metaphysical semantics.

8.3 Further work

In this last section I describe possible direction for further work that extends the setting of this thesis.

Indeterminacy of interpretation

To obtain a complete account of radical interpretation it would be worthwhile to investigate the indeterminacy of interpretation more thoroughly than I do in this thesis. One might try evaluate an account of interpretation on the determinacy requirement with a similar level of formal sophistication as I use to evaluate accounts on the variety requirement.

It should be possible to prove formal results that are complementary to the representation results from Chapter 7. For a given account of interpretation these results would have the following structure:

First, we define a notion of similarity between models, such that we think of two models as being the same as far as the determinacy requirement is concerned if and only if they are similar according to the notion of similarity. Since the many models used in this thesis are also models that are used in modal logic it would be natural to use the notion of bisimilarity from modal logic to capture sameness of models.

Second, we proof a result that establishes conditions on linguistic behaviors such that a linguistic behavior satisfies the conditions if and only if any two models that interpret the linguistic behavior are similar according to the notion of similarity defined in the first step. The conditions on the linguistic behaviors would guarantee that the behavior is rich enough to determine a unique interpreting model. We would need conditions that entail for instance that the subject does not accept only tautologies in all situation or that any two worlds can be separated by the belief or evidence set of the subject in some situation.

It might well be that such results can be proven for different notions of similarity yielding different conditions on linguistic behaviors. In this case there would be a certain variability in the determinacy requirement. Whether an account fulfills the determinacy requirement depends on how much one is willing to weaken the notion of similarity between models and on how much one is willing to strengthen the conditions that guarantee that the linguistic behavior of the subject is rich enough.

Meaning change

As noted at the end of Section 5.3 and again at the end of Section 6.2 it would be good to have a theory of how meanings change across different situations if one uses the disquotational acceptance principle. I see two challenges for developing such a theory.

First, we need to understand when meanings change. If we were to allow meanings to change between any two situations in which we are interpreting the subject then we would be back at the indeterminacy from Section 2.3 that acceptance in a single situation is not enough to determine an interpretation function. This indeterminacy is the reason why we started in Section 2.5 to interpret the subject in different situations with the assumption that her interpretation function in all those situations is the same.

Second, we need to explain how a change in the belief set of the subject can trigger a change in the supervaluation that represents her language. This happens for instance in the example from Section 5.3 where the meaning of a sentence in the language of the subject becomes determinate once she observes how other speakers use the sentences. It might also help to account for the example from Section 6.2 where the subject changes her application of the term "water" upon learning that the molecular structure of water is H_2O.

No such additional theory is required in the case of metasemantic acceptance.

With metasemantic acceptance a change of the meanings that influence the linguistic behavior of the subject is captured as a change of her beliefs about the semantic facts. Hence we can apply the existing theory of belief change to account for the change of meanings. Nevertheless, it would be nice to have some understanding of the prior beliefs, or plausibilities, that humans usually have about the semantic facts. For instance there might be a tendency to adapt our own beliefs about meaning to the usage of the speakers around us.

Interpreting interpreters

An obvious extension of the setting from this thesis is to consider the case of interpreting multiple subjects that interpret each other's linguistic behavior. Thus they form beliefs about each others beliefs and meanings. Such an account would need to employ proper Kripke models to represent beliefs about beliefs and metasemantic models, even in the case of disquotational acceptance, to represent beliefs about meanings. Because beliefs about the beliefs of others play a crucial role in doxastic logic this extension to the multiple subject might be especially important if one wants to use a formal account of interpretation as a foundation for doxastic logic.

A setting that includes multiple subjects could be used to model communication between subjects, which is after all the purpose of language. This might be a challenging problem because interpretation presupposes that the interpreter observes the linguistic behavior of the subject across many different situations whereas communication is most interesting if the speaker informs the hearer about a situation in which the hearer has not been before. To resolve this tension it might be helpful to have an account of the subject's prior beliefs or plausibilities about each others language, such that they are able to communicate without having completely interpreted each other in advance.

In a framework that models beliefs about beliefs and meanings one might also consider subjects that interpret their own linguistic behavior. One could look for conditions on the subject's evidence about her own behavior that guarantee or allow interpretability by a model that satisfies introspection for the subject's beliefs about her own beliefs or meanings. In the case where belief is replaced with knowledge this is the approach taken by Williamson (1994, ch. 8) to argue against introspection and thus allow for higher-order vagueness.

Eligible meanings

In Sections 2.2 and 2.3 I discuss two approaches for eliminating the indeterminacy in the account of Section 2.1, which assumes that the only input for interpretation is the set of sentences that the subject accepts. The first is to assume that we also know the meaning of sentences in the language of the subject prior to interpretation and the other is to assume that we know her beliefs prior to interpretation. In his discussion of Putnam's paradox Lewis (1984, pp. 226–229, but see also 1983, pp. 370–377) advocates another solution to the indeterminacy that he ascribes to Merrill (1980). I think that this solution would be a valuable addition to an account of interpretation. Let me sketch the idea in the following paragraphs.

Lewis addresses Putnam's paradox and not the problem of radical interpretation as it is set up in this thesis and hence he is concerned with the properties and objects that are the referents of expressions in a first-order language and not the propositions that are the meanings in the propositional languages of this thesis. Nevertheless I describe the solution suggested by Lewis in the setting of this thesis.

The solution assumes that the set of all meanings, in our case propositions, is ordered depending on how eligible they are to be the meaning of expression in the language of the subject. Propositions that correspond to natural facts in the world are more eligible than propositions that are just arbitrary sets of possible worlds. For instance the propositions that it is raining is a more eligible meaning than the proposition that it is raining and the lights are on, or, the sun is shining and the man is 1.6 meters tall.

In cases of indeterminacy the eligibility ordering on propositions can help to discriminate between different possible interpretations of the subject's behavior. We choose an interpreting model that maximizes the eligibility of the propositions that the interpretation function of the model assigns to the atomic sentences in the vocabulary of the subject. For example we prefer an interpreting model according to which p means that it is raining to an interpreting model according to which all atomic sentence other than p have the same meaning as in the first model but p means that it is raining and the lights are on, or, the sun is shining and the man is 1.6 meters tall.

It should be possible to make this idea formally precise. An eligibility ordering on propositions could be lifted to an ordering on interpretation functions by defining that I is at least as eligible as I' if for all atomic sentences p the proposition $I(p)$ is at least as eligible as $I'(p)$. Using some of the techniques for lifting relations to the powerset of their carrier one might lift this order even further to an ordering on supervaluations or on belief sets in a metasemantic model.

I think that this idea of ordering the interpreting models according to their eligibility might be helpful to solve a problem that is not the indeterminacy of interpretation. Let me first explain which problem this is and how it is distinct from the indeterminacy of interpretation.

In many cases interpreters do not have enough information about the linguistic behavior of subject to interpret it with a sufficiently unique model. It might be that the interpreters fail to figure out for every sentence whether the subject accepts it or they might fail to interpret the subject in enough different situations to narrow down the possible meanings of her sentences. This problem could be solved by letting the interpreter focus on the most eligible models that interpret some behavior that is compatible with the evidence that the interpreter has about the subject's behavior. In this way the eligibility order over interpreting models functions as a plausibility ordering, in the sense of Chapter 3.3, for interpreters. I have the hope that this idea might help to explain how subjects can communicate facts about situations where they have never been before, which as suggested in the previous subsection might be difficult for an account of interpretation.

The problem that the interpreter might not have enough information about the behavior of the subject is distinct from the problem of indeterminacy of interpretation. The indeterminacy of interpretation only concerns the uncertainty that is left after the interpreter has gathered all necessary information about the behavior of the subject. I am not sure that eligibility orderings are a good solution to an indeterminacy of interpretation. It seems that it rather gives us a way to cope with an indeterminacy instead of resolving it. But this should not keep us from using eligibility orders in cases of uncertainty that does not result from an indeterminacy of interpretation.

Modeling belief

To strengthen the connection between the theory of interpretation and doxastic logic and the theory of belief revision it would be helpful to more thoroughly understand the setting from Section 3.3 in which subject's belief set is given by the most plausible worlds that are compatible with her evidence.

I am only able to characterize the plausibility interpretable behaviors in the case of splitting interpretability and for the most general kind of plausibility orders. One might try to show similar results for complete orders or to some notion of tight interpretability. Example 7.5.4 shows that for plausibility orders supervaluation interpretability does not coincide with splitting interpretability. But I do not understand why this is the case and whether it is desirable or not.

One might also see what happens if we replace the plausibility order with a probability distribution. This would require that we adapts the acceptance principle to probabilistic beliefs. For instance we might say that the subject accepts a sentence if and only if it expresses a proposition that has a probability that is greater than one half. This would give rise to a notion of interpretable behavior that is not a theory because the set of sentences accepted in this way can fail to be closed under conjunctions.

The theory of interpretation and decision theory

A further problem is to clarify the relation between the account of interpretation given in this thesis and decision theory. It might be possible to reduce the account of radical interpretation of this thesis to decision theory. For beliefs reductions of quantitative notions of belief to probabilistic beliefs have been studied by Leitgeb

(2013) and by Lin and Kelly (2012). A reduction of meaning to the subject's desire for communicating certain states of the world is given by the signaling games introduced by Lewis (1969).

Larger fragments of natural language

In this thesis I consider the case in which we have a hypothesis about what expressions in the language of the subject function as the classical propositional connectives or as the necessity modality. One could extend the account such that this hypothesis covers larger fragments of natural languages for which we have a formal semantics.

The first step into this direction is to cover higher-order logics of the kind that are employed by Montague grammar. A challenge for this approach is to explain how the subject forms beliefs about the objects in the domain of the models for such higher-order logics.

Parsing

When we interpret the subject utterances with a hypothesis about what expressions function as the propositional connectives and the necessity modality then this presupposes that we can parse the sentences in the language of the subject to determine the propositional or modal formulas that they correspond to. This assumption becomes even stronger if we extend the account to higher-order logics that uncover the subsentential syntactic structure of sentences. In that case we would need to parse sentences to obtain syntactic trees that assign syntactic categories to the expressions in the language of the subject.

It would be interesting to see whether this parsing of sentences in the language of the subject can be made part of the account of interpretation. The input for interpretation would then contain sentences that are sequences of words in the vocabulary of the subject. An interpreting model would need to provide a classification of these words into syntactic categories.

More refined models of meaning

In this thesis I represent the meaning of sentences with sets of possible worlds as is standard in possible world semantics. The only exception are the two-dimensional meanings from Section 6.4. But It should be possible to adapt the account of interpretation to other, more complex, representations of meaning. Let me mention four extensions in greater detail.

One could try to adapt the formal models such that they can cope with the indexicality of expressions in natural language (see Kaplan 1989). It might already suffice to think of the elements in the domain W as centered possible worlds. In this case one should test on concrete examples whether the two-dimensional supervaluation models with disquotational acceptance and metasemantic models with metasemantic acceptance are already powerful enough to capture the interaction of indexicals with the necessity modality.

One might consider a relativist notion of meaning (see for instance MacFarlane 2014). This probably requires adding standards of assessment to the centering of worlds. It would be interesting to see what kind of linguistic behavior could make use of this parameter and distinguish the relativist view of meaning from a non-indexical contextualist view, as explained by MacFarlane (2009). Maybe we need to include a basic linguistic act of assessing the utterances of other subjects into the notion of a linguistic behavior.

It would be interesting to develop an account of interpretation for a system of dynamic semantics (see for instance Groenendijk, Stokhof, and Veltman 1996). It seems that in such systems the relation between interpretational meaning and compositional meaning is much more complex than described in the previous section. For instance the compositional meaning of anaphora and epistemic modals depends on a local, partially metasemantic, information state that is associated to the syntactic context in which the expression is embedded. I am not sure that the distinction between disquotational and metasemantic acceptance still makes sense in such a setting. Another issue is whether the representation of the mental state of the subject should be adapted to the dynamic framework, as suggested for instance by Kamp (1990).

One could also develop and account of interpretation for an inquisitive notion of meaning (see for instance Ciardelli, Groenendijk, and Roelofsen 2013). This might require to adapt the notion of a linguistic behavior to include the posing of questions and to enrich the representation of the mental state to include the issues that the agent entertains. The latter has already been done by Ciardelli and Roelofsen (2015) in the context of epistemic logic.

What are semantic facts?

A central conceptual question that I do not address in this thesis is the question what the semantic facts are supposed to be. I think that different answers to this question influence what account of interpretation we want to use and vice versa different accounts of interpretation suggest different answers to this question.

One basic issue is whether semantic facts are equivalent to or constrained by facts that are independent from the linguistic behavior of the subject. I hastily dismiss this possibility at the end of Section 2.2 to move on to an account on which the interpretation function is determined by interpretation of linguistic behavior. But this matter deserves more careful consideration. Especially it would be interesting to see what happens if we reintroduce the idea that some semantic facts are determined by facts that are independent from the subject's behavior to an account of interpretation that presupposes knowledge about some of the subject's beliefs.

Consider the semantic fact that the sentence p means that Q and choose a basic fact F that is independent of the subject's linguistic behavior. Suppose that F constrains the semantic fact that p means that Q in such a way that whenever F is the case then p expresses the proposition Q. If we use disquotational acceptance this would mean that whenever F is the case in the actual world then we have to interpret the subject with an interpretation function that maps p to Q. It would be interesting characterize the condition that this imposes on the interpretable behaviors. I have the impression that for almost all choices of F this condition would impose quite strong constraints on the interpretable behaviors.

With metasemantic acceptance the situation is different. Suppose that the relation that whenever F is the case then p expresses the proposition Q is required to hold at all possible world in the domain of an interpreting metasemantic model. This seems to impose the condition on the interpretable behaviors that whenever the subject believes or infers from experience that F then she has to accept p if and only if she believes Q. It would be nice to make this condition formally precise and evaluate its plausibility.

Also in the case where semantic facts are determined exclusively by linguistic behavior their status is not entirely clear. With disquotational acceptance it seems that they would end up being about the subject's dispositions for behavior, similar to the facts about utility in decision theory. With metasemantic acceptance we would have that believing a certain semantic fact would amount to being disposed to a certain kind of linguistic behavior. It seems that in this case semantic facts play a role similar to the normative facts in an expressivist theory. Such a view of semantic facts, although for a different conception of meaning, has been advocated by Gibbard (2012).

Bibliography

Baltag, Alexandru and Sonja Smets (2006). “Conditional Doxastic Models: A Qualitative Approach to Dynamic Belief Revision.” In: *Electronic Notes in Theoretical Computer Science* 165, pp. 5–21.

Blackburn, Patrick, Maarten de Rijke, and Yde Venema (2002). *Modal Logic*. Cambridge University Press.

Burge, Tyler (1979). “Individualism and the Mental.” In: *Midwest Studies in Philosophy* 4.1, pp. 73–122.

— (1986). “Intellectual Norms and Foundations of Mind.” In: *Journal of Philosophy* 83.12, pp. 697–720.

Carnap, Rudolf (1947). *Meaning and Necessity*. University of Chicago Press.

Chalmers, David J. (2002). “The Components of Content.” In: *Philosophy of Mind: Classical and Contemporary Readings*. Ed. by David J. Chalmers. Oxford University Press.

— (2006). “The Foundations of Two-Dimensional Semantics.” In: *Two-Dimensional Semantics: Foundations and Applications*. Ed. by Manuel García-Carpintero and Josep Macia. Oxford University Press, pp. 55–140.

Ciardelli, Ivano, Jeroen Groenendijk, and Floris Roelofsen (2013). “Inquisitive Semantics: A New Notion of Meaning.” In: *Language and Linguistics Compass* 7.9, pp. 459–476.

Ciardelli, Ivano and Floris Roelofsen (2015). “Inquisitive dynamic epistemic logic.” In: *Synthese* 192.6, pp. 1643–1687.

Davidson, Donald (1973). “Radical Interpretation.” In: *Dialectica* 27.1. page references are to (Davidson 2001, pp. 125–139), pp. 314–328.

— (1974). “Belief and the Basis of Meaning.” In: *Synthese* 27.3/4. page references are to (Davidson 2001, pp. 141–154), pp. 309–323.

— (1975). “Thought and Talk.” In: *Mind and Language*. Ed. by Samuel D. Guttenplan. Clarendon Press, 7–23.

— (1979). “The Inscrutability of Reference.” In: *Southwestern Journal of Philosophy* 10.2, pp. 7–19.

Davidson, Donald (1980). “Toward a Unified Theory of Meaning and Action.” In: *Grazer Philosophische Studien* 11, pp. 1–12.

— (1992). “Three Varieties of Knowledge.” In: *A. J. Ayer Memorial Essays*. Ed. by A. Phillips Griffiths. Cambridge University Press, pp. 153–166.

— (2001). *Inquiries Into Truth and Interpretation*. Oxford University Press.

Dretske, Fred (1981). *Knowledge and the Flow of Information*. MIT Press.

Fagin, Ronald et al. (2003). *Reasoning About Knowledge*. MIT Press.

Fine, Kit (1975). “Vagueness, Truth and Logic.” In: *Synthese* 30.3-4, pp. 265–300.

Fritz, Peter (2011). “Matrices and modalities: On the logic of two-dimensional semantics.” Master’s thesis. ILLC, University of Amsterdam.

Gamut, L.T.F. (1991). *Intensional Logic and Logical Grammar*. Vol. 2. Logic, Language, and Meaning. University of Chicago Press.

Gärdenfors, Peter (1973). “On the extensions of S5.” In: *Notre Dame Journal of Formal Logic* 14.2, pp. 277–280.

Gibbard, Allan (2003). *Thinking How to Live*. Harvard University Press.

— (2012). *Meaning and Normativity*. Oxford University Press.

Groenendijk, Joroen, Martin Stokhof, and Frank Veltman (1996). “Coreference and Modality.” In: *The Handbook of Contemporary Semantic Theory*. Ed. by Shalom Lappin. Blackwell, pp. 179–216.

Grove, Adam (1988). “Two Modellings for Theory Change.” In: *Journal of Philosophical Logic* 17.2, pp. 157–170.

Halpern, Joseph Y. and Riccardo Pucella (2011). “Dealing with logical omniscience: Expressiveness and pragmatics.” In: *Artificial Intelligence* 175.1, pp. 220 –235.

Heim, Irene and Angelika Kratzer (1998). *Semantics in Generative Grammar*. Blackwell.

Hintikka, Jaakko (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press.

Jeffrey, Richard C. (1983). *The Logic of Decision*. University of Chicago Press.

Kamp, Hans (1971). “Formal Properties of ‘Now’.” In: *Theoria* 37, pp. 227–274.

— (1975). “Two Theories about Adjectives.” In: *Formal Semantics of Natural Languages*. Ed. by Edward L. Keenan. Cambridge University Press, pp. 123–155.

— (1990). “Prolegomena to a Structural Account of Belief and Other Attitudes.” In: *Propositional Attitudes: The Role of Content in Logic, Language, and Mind*. Ed. by C. Anthony Anderson and Joseph Owens. CSLI, pp. 27–90.

Kaplan, David (1989). “Demonstratives.” In: *Themes From Kaplan*. Ed. by Joseph Almog et al. Oxford University Press, pp. 481–563.

Kraus, Sarit, Daniel Lehmann, and Menachem Magidor (1990). “Nonmonotonic Reasoning, Preferential Models and Cumulative Logics.” In: *Artificial Intelligence* 44.1-2, pp. 167–207.

Kripke, Saul A. (1979). “A Puzzle About Belief.” In: *Meaning and Use*. Ed. by Avishai Margalit. Reidel, pp. 239–283.

— (1980). *Naming and Necessity.* Harvard University Press.

Leitgeb, Hannes (2013). "Reducing Belief Simpliciter to Degrees of Belief." In: *Annals of Pure and Applied Logic* 164.12, pp. 1338–1389.

Lewis, David (1969). *Convention: A Philosophical Study.* Harvard University Press.

— (1973). *Counterfactuals.* Blackwell.

— (1974). "Radical Interpretation." In: *Synthese* 27, pp. 331–344.

— (1975). "Languages and Language." In: *Minnesota Studies in the Philosophy of Science.* Ed. by Keith Gunderson. Vol. 7. University of Minnesota Press, pp. 3–35.

— (1979). "Attitudes de Dicto and de Se." In: *Philosophical Review* 88.4, pp. 513–543.

— (1980). "Index, Context, and Content." In: *Philosophy and Grammar.* Ed. by Stig Kanger and Sven Öhman. Reidel, pp. 79–100.

— (1983). "New Work for a Theory of Universals." In: *Australasian Journal of Philosophy* 61.4, pp. 343–377.

— (1984). "Putnam's Paradox." In: *Australasian Journal of Philosophy* 62.3, pp. 221–236.

— (1996). "Elusive Knowledge." In: *Australasian Journal of Philosophy* 74.4, pp. 549–567.

Lin, Hanti and Kevin T. Kelly (2012). "Propositional Reasoning that Tracks Probabilistic Reasoning." In: *Journal of Philosophical Logic* 41.6, pp. 957–981.

MacFarlane, John (2009). "Nonindexical Contextualism." In: *Synthese* 166.2, pp. 231–250.

— (2014). *Assessment Sensitivity: Relative Truth and its Applications.* Oxford University Press.

Marti, Johannes and Riccardo Pinosio (2016). "A Game Semantics for System P." forthcoming in *Studia Logica*, available as an ILLC prepublication PP-2016-10.

McCarthy, Timothy (2002). *Radical Interpretation and Indeterminacy.* Oxford University Press.

Menzel, Christopher (2015). "Possible Worlds." In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Spring 2015.

Merrill, G. H. (1980). "The Model-Theoretic Argument against Realism." In: *Philosophy of Science* 47.1, pp. 69–81.

Putnam, Hilary (1975). "The Meaning of 'Meaning'." In: *Minnesota Studies in the Philosophy of Science* 7, pp. 131–193.

— (1981). *Reason, Truth, and History.* Cambridge University Press.

Quine, W. V. (1960). *Word and Object.* MIT Press.

— (1968). "Ontological Relativity." In: *Journal of Philosophy* 65.7, pp. 185–212.

Rayo, Augustin (2013). *The Construction of Logical Space.* Oxford University Press.

Rott, Hans (2014). “Four Floors for the Theory of Theory Change: The Case of Imperfect Discrimination.” In: *Logics in Artificial Intelligence.* Ed. by Eduardo Fermé and João Leite. Vol. 8761. Lecture Notes in Computer Science, pp. 368–382.

Savage, Leonard J. (1972). *The Foundations of Statistics.* Dover Publications.

Schroeder, Mark (2008). “What is the Frege-Geach Problem?” In: *Philosophy Compass* 3.4, pp. 703–720.

Scroggs, Schiller Joe (1951). “Extensions of the Lewis System S5.” In: *The Journal of Symbolic Logic* 16.2, pp. 112–120.

Soames, Scott (2005). *Reference and Description: The Case Against Two-Dimensionalism.* Princeton University Press.

Stalnaker, Robert C. (1978). “Assertion.” In: *Syntax and Semantics.* Ed. by Peter Cole. Vol. 9. New York Academic Press, pp. 315–332.

— (1984). *Inquiry.* Cambridge University Press.

— (1990). “Narrow Content.” In: *Propositional Attitudes: The Role of Content in Logic, Language, and Mind.* Ed. by C. Anthony Anderson and Joseph Owens. CSLI, pp. 131–145.

— (1999). “The Problem of Logical Omniscience, II.” In: *Context and Contend.* Oxford University Press, pp. 255–273.

— (2001). “On Considering a Possible World as Actual.” In: *Proceedings of the Aristotelian Society* 75, pp. 141–156.

— (2004). “Assertion Revisited: On the Interpretation of Two-Dimensional Modal Semantics.” In: *Philosophical Studies* 118.1-2, pp. 299–322.

Stokhof, Martin (2014). “Arguing About Dynamic Meaning.” In: *Johan van Benthem on Logic and Information Dynamics.* Ed. by Alexandru Baltag and Sonja Smets. Springer, pp. 749–764.

Van Ditmarsch, Hans, Wiebe van der Hoek, and Barteld Kooi (2007). *Dynamic Epistemic Logic.* Springer.

Van Fraassen, Bas C. (1966). “Singular Terms, Truth-Value Gaps, and Free Logic.” In: *Journal of Philosophy* 63.17, pp. 481–495.

— (1977). “The Only Necessity is Verbal Necessity.” In: *Journal of Philosophy* 74.2, pp. 71–85.

— (1979). “Propositional Attitudes in Weak Pragmatics.” In: *Studia Logica* 38.4, pp. 365–374.

Veltman, Frank (1985). “Logics for Conditionals.” PhD thesis. University of Amsterdam.

Williamson, Timothy (1994). *Vagueness.* Routledge.

Yalcin, Seth (2014). “Semantics and Metasemantics in the Context of Generative Grammar.” In: *Metasemantics: New Essays on the Foundations of Meaning.* Ed. by Alexis Burgess and Brett Sherman. Oxford University Press, pp. 17–54.

Index

Samenvatting

Interpretatie van talig gedrag met behulp van mogelijke-werelden-modellen

In dit proefschrift ontwikkel ik een benadering van radicale interpretatie, daarbij gebruikmakend van het raamwerk van mogelijke werelden om geloofstoesstanden en betekenis te modelleren. Ik bewijs representatiestellingen die laten zien dat als het talige gedrag van een subject aan bepaalde voorwaarden voldoet, dit gedrag kan worden gezien als voortkomend uit een bepaald type mogelijke-werelden-model van geloof en betekenis. De bewezen stellingen zijn analoog aan representatietheorema's in de beslistheorie, die laten zien dat, indien het keuzegedrag van een subject aan bepaalde voorwaarden voldoet, dit gedrag voortkomt uit het maximaliseren van de nutsverwachting gegeven een bepaalde subjectieve waarschijnlijkheids- en nutsfunctie.

Ik beschouw verschillende instellingen, die variëren in de details van het modelleren. De resulterende benaderingen bestaan alle uit de volgende vijf stappen:

Ten eerste, een definitie van de gebruikte mogelijke-werelden-modellen. In simpele gevallen bevat dit model een verzameling mogelijke werelden om de geloofstoestand van een subject te representeren en een functie van zinnen naar verzamelingen van werelden om betekenissen te representeren.

Ten tweede, een definitie van een notie van talig gedrag. In het basisgeval neem ik als talig gedrag een verzameling zinnen in de taal van het subject, die gezien wordt als de verzameling zinnen die het subject in een zekere situatie accepteert.

Ten derde, een definitie van het talige gedrag dat gegenereerd wordt door een zeker mogelijke-werelden-model. Deze definitie is een formalisatie van het idee dat het subject in een situatie een zin accepteert, indien deze een propositie uitdrukt die volgens de geloofstoestand van het subject waar is in die situatie.

Ten vierde, een specificatie van verdere aannames betreffende het subject. Hieronder vallen aannames over de betekenis van bepaalde uitdrukkingen in de taal van het subject, zoals de propositionele connectieven, en aannames over wat de spreker gelooft in bepaalde situaties, zoals geloof dat het subject ontleent aan de zintuiglijke waarneming.

Ten vijfde, een representatietheorema dat noodzakelijke en voldoende voorwaarden geeft, waaronder een talig gedrag precies dat gedrag is dat gegenereerd wordt door een model dat aan de voorwaarden in de vierde stap voldoet.

In het eerste deel van het proefschrift ontwikkel ik een elementaire benadering van interpretatie, die verderop in het proefschrift uitgebreid zal worden, en beschouw ik verschillende mogelijkheden om geloofstoestanden te modelleren. De meest verfijnde benadering die ik behandel, maakt gebruik van een ordening van mogelijke werelden op basis van hun plausibiliteit, om te modelleren hoe het subject haar geloof bijstelt in het licht van nieuwe informatie.

In het tweede deel van het proefschrift verken ik verschillende manieren om betekenissen te modelleren. Ik onderscheid een disquotationeel van een metasemantisch acceptatieprincipe. Volgens het disquotationele acceptatieprincipe accepteert het subject een zin dan, en slechts dan, als volgens de semantische feiten in de actuele wereld de zin een propositie uitdrukt die het subject gelooft. Volgens het metasemantische acceptatieprincipe accepteert het subject zinnen dan, en slechts dan, als het subject gelooft dat de zin een ware propositie uitdrukt.

Het onderscheid tussen het disquotationele en metasemantische acceptatieprincipe heeft een cruciale invloed op het formele model van betekenis. Ik demonstreer dit aan de hand van twee concrete problemen voor een theorie van interpretatie waarvan de oplossing afhangt van het gekozen acceptatieprincipe.

Het eerste probleem is het verklaren van situaties waarin het subject, voor een bepaalde zin, ondanks volledige kennis over alle relevante feiten, noch die zin noch zijn negatie accepteert. Volgens het metasemantische acceptatieprincipe is het subject in dergelijke situaties onzeker over de semantische feiten die de betekenis van de zin bepalen. Volgens het disquotationele acceptatieprincipe zijn zulke situaties alleen te verklaren met behulp van een notie van betekenis volgens welke de waarheidswaarde van een zin in een mogelijke wereld onbepaald kan zijn.

Het tweede is het probleem van noodzakelijkheid a posteriori, dat zich voordoet wanneer de acceptatie van zinnen met een noodzakelijkheidsmodaliteit verandert, voor het subject, wanneer zij nieuwe informatie over de wereld verkrijgt. Ik beargumenteer dat het onderscheid tussen het disquotationele en metasemantische acceptatieprincipe correspondeert met het onderscheid tussen de epistemische en metasemantische interpretatie van twee-dimensionale semantiek.

Summary

Interpreting linguistic behavior with possible world models

Interpreting Linguistic Behavior with Possible World Models

Johannes Marti

In this thesis I develop an account of radical interpretation using the possible world framework to model beliefs and meanings. I prove representation results which show that if the linguistic behavior of some subject satisfies certain conditions then it can be taken to arise from a certain type of possible world model for belief and meaning. These results are analogous to representation theorems in decision theory which show that if the choice behavior of some subject satisfies certain conditions then it arises from expected utility maximization with respect to some subjective probability and utility functions.

I consider multiple accounts that differ in the details of the modeling. They have in common that they all consist of the following five steps:

First, a definition of the possible world models that are used. In simple cases they contain a set of possible worlds to represent the belief state of some subject and a function mapping sentences to sets of worlds to represent meanings.

Second, a definition of some notion of linguistic behavior. In the basic case I take a linguistic behavior to be a set of sentences in the language of the subject which is thought of as the set of sentences the subject accepts in some situation.

Third, a definition of the linguistic behavior generated by some possible world model. This definition is a formalization of the idea that the subject accepts some sentence in some situation if the sentence expresses a proposition that the subject believes in that situation.

Forth, a specification of further assumptions that we are making about the subject. This includes assumptions about the meaning of some expressions in the language of the subject, such as for instance the propositional connectives, or

assumptions about some of the subject's beliefs in certain situations, such as for instance the beliefs that the subject obtains from perception.

Fifth, a representation results that gives necessary and sufficient conditions for a linguistic behavior to be the behavior generated by some model that satisfies the additional assumptions made in the forth step.

In the first part of the thesis I develop the basic account of interpretation, which is extended later, and consider different possibilities for modeling beliefs. The most refined account that I discuss uses a plausibility order over possible worlds to model how the subject revises her beliefs in light of new evidence.

In the second part of the thesis I explore different possibilities for modeling meanings. I distinguish between a disquotational and a metasemantic acceptance principle. According to the disquotational acceptance principle the subject accepts a sentence if and only if according to the semantic facts that obtain at the actual world the sentence expresses a proposition that the subject believes. According to the metasemantic acceptance principle the subject accepts a sentences if and only if the subject believes that the sentence expresses a true proposition.

The distinction between disquotational and metasemantic acceptance crucially influences the formal model of meaning. I demonstrate this by discussing two concrete problems for a theory of interpretation that are solved differently depending on which version of the acceptance principle we choose.

The first is the problem of accounting for situations in which the subject has complete knowledge about all the relevant facts but still does not accept some sentence nor accepts its negation. On the metasemantic account of acceptance such cases are situations in which the subject is uncertain about the semantic facts that determine the meaning of the sentence. With a disquotational account of acceptance one can only account for such cases if one employs a notion of meaning according to which the truth value of some sentence at some possible world can remain indeterminate.

The second is the problem of necessity a posteriori that arises when the subject's acceptance of sentences containing a necessity modality changes as she obtains new beliefs about the world. I argue that the distinction between disquotational and metasemantic acceptance corresponds to the distinction between the epistemic and the metasemantic interpretation of two-dimensional semantics.

Titles in the ILLC Dissertation Series:

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models