The Dissertation Committee for Georgui Kirillovich Bronnikov
certifies that this is the approved version of the following dissertation:

# Representation of Inference in the Natural Language

Committee:

Nicholas Asher, Supervisor

David Beaver

Daniel A. Bonevac

Joshua Dever

Martina Faller

Robert C. Koons

# Representation of Inference in the Natural Language

by

**Georgui Kirillovich Bronnikov, B.A.**

**DISSERTATION**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**DOCTOR OF PHILOSOPHY**

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2011

# Acknowledgments

I got my undergraduate degree at the Theoretical and Applied Linguistics Department (FTiPL) of the Russian State University for Humanities (RGGU). The department had been established due to the efforts of its first head, Alexander Nikolaevich Barulin, and he played a major part in creating the joyful and stimulating athomsphere of science, learning and cooperation. Among the professors that I had the privilege to learn from were Vera Isaakovna Podlesskaya, Zoya Mikhailovna Shalyapina, Ekaterina Vladimirovna Rakhilina and Yakov Georgievich Testelec. I got my first acquaintance with the field of formal semantics in the classes taught by Barbara Partee, and she had a profound influence on the way I view linguistics.

At the University of Texas, Nick Asher was my supervisor. Without his help and encouragement this dissertation would have probably never been started, much less brought to the defense stage. In my work, I got valuable advice from David Beaver, Ian Buchanan and Hans Kamp. Martina Faller, as the external member of the committee, also provided extremely helpful comments on the draft. I would especially like to thank Josh Dever for his series of logic courses. Back in Moscow, Lev Beklemishev helped me work on the logic part of the project.

I enjoyed the friendship of my fellow students Tristan Johnson, Malte Willer, Anita Huang, Guha Krishnamurthi, Julie Hunter, Hsiang-Yun Chen, Mina Chen, Elias Ponvert, Hilaria Cruz and Elena Liskova.

Jill Glenn and Sally Jackman, Graduate Coordinators of the Philosophy

department, did everything in their power to help me understand the customs and rituals of academic life in a foreign country.

And of course, I could always rely on my family — my parents Kirill Bronnikov and Elena Bronnikova, my sisters Natasha and Dasha, my wife Maria Rubinstein and my daughter Liza. Thank you for tolerating my long periods of absence!

# Representation of Inference in the Natural Language

Publication No. ____________________

Georgui Kirillovich Bronnikov, Ph.D.
The University of Texas at Austin, 2011

Supervisor: Nicholas Asher

The purpose of this work is to investigate how processes of inference are reflected in the grammar of the natural language. I consider a range of phenomena which call for a representational theory of mind and thought. These constructions display a certain regularity in their truth conditions, but the regularity does not extend to closure under arbitrary logical entailment. I develop a logic that allows me to speak formally about classes of inferences. This logic is then applied to analysis of indirect speech, belief reports, evidentials (with special attention to Bulgarian) and clarity assertions.

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Formal Semantics vs the Representational Theory of Mind

Formal, or model-theoretic, semantics is arguably the most successful theory of meaning for natural language. Sentence meanings are identified with their truth conditions. In order to derive compositionally the infinite set of sentence truth conditions from a finite set of primitive expressions, the primitive expressions get assigned as meanings set-theoretic constructions built on top of truth-values (type `t`), real-world objects (type `e`) and possible worlds (type `s`). Formal semantics as such makes no claims concerning the mechanisms that allow us humans to grasp these kinds of meanings.[1] Clearly, sets of possible worlds cannot be directly stored in the brain (unless we are speaking of some very restricted finite sets).

On the other hand, at least a seriously taken viewpoint in the philosophy of mind is the Representational Theory of Mind (RTM) (Fodor, 1975).

[1] The characterization of formal semantics given here does not take into account the Discourse Representation Theory (DRT) group of views, where, at a minimum, the notion of a sentence meaning is changed (to a so-called context change potential) and a new basic type is added (discourse referents or "pegs"). More significantly for the purposes of this investigation, some proponents of DRT (Asher, 1986; Kamp, 1990; Kamp et al., 2005) argue for a representational account of belief, and certain theories (such as the account of presupposition of van der Sandt (1992)) crucially rely on using representations in computing meanings.

This theory maintains that (at least part of) our cognitive activity consists in manipulating sentence-like representations in our brains. Believing a sentence to be true, to a first approximation, amounts to having a token of this sentence, translated into some internal language, in the appropriate region of our minds (often called “belief box”). Sentences in this language are somehow semantically interpreted, so that we can determine their truth conditions.

A natural component of the Representational Theory of Mind is the Representational Theory of Thinking, stating that our logical abilities consist (at least in part) in building new internal-language representations on the basis of old ones, in a manner similar to inference in logics.

The viewpoint of RTM is in principle compatible with the viewpoint of model-theoretic semantics. It is conceivable that the proper semantics of the natural language is provided by set-theoretic structures, even though thoughts are stored in the mind in sentence-like form. After all, this is how most kinds of logics work: their semantics is based on set-theoretic models while mathematicians build proofs using formulas and inference rules.

The meaning of our language’s sentences could be adequately characterizable by their truth conditions, and sentence-like representations in our minds could be a way of getting at those truth conditions.

The goal of this thesis is to show that such a picture is insufficient; that there exist constructions in natural language whose proper analysis crucially involves the Representational Theory of Mind and sentential representations in the minds of the speakers. Such constructions present language users as reasoners who perform inferences on their internal representations; which classes of inferences are taken into account depends on the particular construction.

## 1.2 Plan of the thesis

In order to make my viewpoint detailed enough for it to have any predictive power, I start off in Chapter 2 by developing a formal logical system that I use in later chapters to couch my analyses in. The logic developed in this chapter is actually too restricted to capture any of the natural language constructions I am going to talk about (in particular, the only variety of the agent's internal language mentioned there is propositional logic; the rules employed have a very simple structure and are non-defeasible), but on one hand, it is simple enough to be rigorously analyzed, and on the other hand, rich enough to serve as a base for extensions that are used for natural language phenomena (different extensions are used for different constructions).

After laying the foundation in Chapter 2, I turn to individual language constructions that show dependence on human inferential capabilities. First, Chapter 3 takes on indirect speech reports. The range of inferential transformations allowed in these reports is narrower than in other constructions I talk about, and this is the shortest of the linguistic chapters. At the same time it serves to demonstrate the overall schema that I use in investigating each group of phenomena.

Chapter 4 talks about belief ascriptions. Of all the constructions studied in this work, belief and belief reports have received the most attention from philosophers. So, unsurprisingly, a large part of that chapter is devoted to the discussion of various points of view, a number of which are very close to my own.

Chapter 5 is concerned with a more exotic topic, morphologically expressed evidentiality. Bulgarian serves here as a case study. Again, in order to even start discussing the issues that worry me, I have to first introduce the

Bulgarian verbal categories in question and the existing research. The problem I am most concerned with are the limits of applicability for the hearsay marker (Renarrated) and the inference marker (Conclusive).

Finally, Chapter 6 deals with clarity assertions. The so called "paradox of clarity" was introduced by Barker and Taranto (2003). I argue against their solution of the paradox and embrace a theory they explicitly reject, the missing inference theory. This work can also be compared to von Fintel and Gillies (2009a), dealing with epistemic *must* in English.

A bit of terminology is in order here. Several constructions under consideration — indirect speech, hearsay evidentials and belief reports when based on believer's own assertions, — involve a *primary utterance* $p$ by a *primary speaker* $A$ and a *secondary utterance* $q$ made by a *secondary speaker* $B$ whose truth conditions depend on the contents of the primary one. For example, when $A$ says:

(1) A: Horses eat oats and hay.

then $B$ later, on the basis of this, can make a secondary utterance:

(2) B: $A$ said that horses eat oats and hay.

For each particular construction, the focus of my investigation will be the relation between the primary and the secondary utterances; namely, what is the range of primary utterances $p$ that can justify a given secondary utterance $q$? And conversely, given a primary utterance $p$, which (true) secondary utterances $q$ can it give rise to?

In each of the linguistic chapters, there is a section towards the end where I try to build an analysis using the formal apparatus of the logic chapter. The formal language of Chapter 2 usually proves insufficient to capture the intricacies of the natural language constructions. In each case, I propose extensions to the formalism. The properties of such extended systems are *not* studied as rigorously as those of the basic system.

Readers whose main interests lie on the linguistic side of my endeavour can safely ignore both Chapter 2 and the formal sections of the later chapters. My linguistic conclusions are largely independent from the logical language I use to make them precise.

## 1.3 Human inference systems

It should be noted that even though in Chapter 2 I concentrate on the case of an agent who uses Natural Deduction as his inference system, I don't expect that the system employed by actual human agents is very similar. In particular, there is no reason to expect that this system will be

- minimal, containing no redundant rules;
- monotonic.

Even when a certain general rule is available to an agent, this may not prevent the same agent from having a number of its more particular instances written into the inference system as separate rules. These particular rules may happen to be more easily accessible than the general one, and a natural language construction may allow use of such rules while prohibiting the general one. An example can be seen on p. 80, in the discussion of the subtyping rule.

Furthermore, many rules employed by human reasoners are heuristic, they don't guarantee the truth of their conclusion with absolute certainty.

This leads to effects like the Paradox of Clarity (Chapter 6).

## 1.4 Limitations of my work

The task I set out to handle — proving that inferential processes deserve a place in proper semantic theory, — occupies nearly all of my attention. This is why I deliberately refrain from handling some of the related questions.

First of all, I never justify my use of RTM; I just take it for granted. The only place where I mention any alternative theories is in §4.1, and the discussion there is limited to the topic at hand, i. e. belief and belief reports.

Second, some of the constructions I work with (especially indirect speech and belief reports) are among the most familiar material in the philosophy of language. However my purposes are different from those of most authors writing about those topics, so I make the following assumptions about my agents:

- they are not confused about identities and they correctly use their general terms, including natural kind terms;
- their language is free of indexicals.

This makes the material I study uninteresting for many philosophers of language; my agents are not confused in ways that help them in their studies. My agents are, however, imperfect in a different way: they have limited logical abilities.

Third, truth conditions for every construction I discuss are vague. Certain uses of the construction are definitely true, certain uses are definitely false, but some just feel uneasy. The only way I can see of dealing with this vagueness is to pretend it does not exist; therefore I present language as being more systematic than it really is. My general strategy in each case is to take the narrowest possible stance: everything suspicious is declared false. This causes my rules to undergenerate; lots of existing usage will be called erroneous, sloppy application of rules. One case where this strategy fails is the Bulgarian data; not being a speaker of the language, I have to take the judgments of my consultants at face value.

## 1.5 Neutrality

Even though I rely heavily on RTM in my work, I wish to remain neutral with respect to many of important statements defended by some of its proponents. In particular:

- I don't claim that RTM is even true. All I care about is that some form of RTM is built into the linguistic picture of the world, the "natural language metaphysics". Pieces of information about the world built into the language itself are known to be sometimes wrong — we talk about the rising of the sun, but it is very hard, if possible at all, to give an analysis of the predicate *rise* that would make this statement literally true. In a similar way, the best description of our cognitive life may involve neural nets with no recognizable syntactic structure, or something else entirely; our common-sense psychology presented by the language may be so wrong that there is no way to make sense of it. What I set out to demonstrate is that the proper presentation of the *linguistic* picture unavoidably includes RTM.[2]

[2] The possibility discussed here is close to the position advocated in (Stich, 1983).

- It does not matter to me whether the internal language, so called Mentalese, is universal; whether it is the same across all of our species or even among speakers of the same natural language. If it is not, I only need personal varieties of Mentalese to be broadly translatable into each other; I will also claim that our linguistic picture includes a scale, a measure of difficulty among various inferences agents can perform.

- It is beyond my sphere of interest how the primitive symbols of Mentalese acquire their meanings and what those meanings consist in (they should at least be sufficient to determine truth conditions of complete formulas in the belief box). I expect that some causal story should eventually be told about acquisition of these meanings; possibly some of the causes in the story operate over the lifetime of our species rather than an individual speaker; the meanings dependent on those will be innate. In any case, nothing in this work depends on any of that.

- In my representations of Mentalese sentences I try to be as close as possible to first-order logic, but this is just for the purposes of presentation. I don't know what the actual syntax of Mentalese looks like; all I hope is that both the formulas I use and the inference rules I ascribe to my agents are translatable into the actual mental code.

Another simplification I make is that I assume tokens of internal language to form separate linearizable sentences. I don't see any reason for the mental code to consist of linear (or even tree-like) pieces. I expect chunks of this code to be more like data in computer memory, with lots of pointers and massive structure sharing. But sentences and formulas are much easier to talk about in linear text, so this is what I use.

# Chapter 2

# Syntactic dynamic doxastic logic

Consider an agent whose state contains a set of formulas in some internal language —his beliefs[1], and who possesses a rule-based inference system (such as natural deduction). By applying rules, the agent is able to add their conclusions to his own state and thus transition into new states. The basic idea is to use the apparatus of dynamic logic to describe these transitions.

In this chapter, I consider two related formal implementations of this basic idea. In Section 2.1, an approach with an agent using a Fitch-style deduction system is given. In Section 2.2 the agent is taken to be using Gentzen-style proofs. The resulting logic is strictly more expressive, but, in the general case, undecidable. Section 2.3 compares my logics to other related systems. Section 2.4 discusses some possible extensions[2].

[1] Here I am using the word 'belief' in a technical sense; no claim is made that this sense corresponds to the meaning of the natural-language verb 'to believe'. Belief ascriptions in natural language are studied in Chapter 4.

[2] In the course of this chapter, I use the letters $p, q$ for sentential variables of the agent's internal language; $a, b$ for metavariables, $\zeta, \eta, \theta, \xi$ for arbitrary formulas of the internal language, as well as for formula patterns, $\alpha, \beta$ for arbitrary actions, $\mathrm{R}, \mathrm{Q}$ for rules of the agent's logic/elementary actions in the external language, $x, y$ for CFDL action variables, $\phi, \chi, \psi$ for formulas of the external language, $s, t$ for states in the model.

## 2.1 Logic for a Fitch-style reasoning agent

Let us start to make the idea more specific. The first step is to assume that the agent's internal language $\mathcal{L}$ is context-free with no ambiguities (or, equivalently, that its formulas are stored in the agent's memory as syntactic trees), and that formulas can have other formulas as constituents.

I am also going to use formula patterns. Let us have a countable set of metavariables $a, b, \ldots, a_1, a_2, \ldots$. A *formula pattern* is a formula where zero or more subformulas are replaced by metavariables. Given a formula pattern $\eta$ and a substitution $\sigma$ that takes metavariables to formulas, with $\mathrm{dom}(\sigma)$ including all the metavariables in $\eta$, $\sigma\eta$ is called an instance of the pattern.

We shall also, for now, assume that the agent uses a Fitch-style calculus for his proof system, and so the rules he employs are of two types: simple rules and subproof rules. Simple rules have the form

$$\frac{\zeta_1 \qquad \zeta_2 \qquad \ldots \qquad \zeta_n}{\eta}\ \mathrm{R}$$

where R is the name of the rule, and $\zeta_i$ and $\eta$ are formula patterns. Subproof rules have the form

$$\begin{array}{c} \zeta \\ \ldots \\ \dfrac{\eta}{\theta}\ \mathrm{Q} \end{array}$$

where Q is the name of the rule, and $\zeta$, $\eta$ and $\theta$ are formula patterns.

Having thus established what the agent's internal language and proof system are, we turn to the task of specifying the external language $F_{\mathcal{L}}$ we will use to describe the agent's states.

For an internal-language formula $\eta$, we shall use the external-language formula $\mathbf{B}\eta$ to denote that the agent has $\eta$ in his belief state.

We will use rule names as elementary actions in dynamic logic. Here the question of granularity becomes evident: how finely do we distinguish between rules? One solution is to treat every instance of a rule as a separate action. In that case, however, the agent can only use this action at most once to change his state: after that, the rule's conclusion will already be a part of his beliefs. The resulting system will be simpler to analyze, but it won't be possible to indicate that a certain general rule, or a certain combination of such rules, is sufficient to derive a certain conclusion when applied repeatedly. Also, from the formal point of view, the system I propose has a finite set of elementary actions, while if we consider each instance of a rule a separate action, this number will be infinite.

We are ready now to define the syntax and semantics of our external language. First, the syntax.

`Definition` **2.1.1** *The set of* actions *is the smallest set produced by the following rules:*

1. *Each rule* R *is an action;*
2. *If $\alpha$ and $\beta$ are actions, then $\alpha \cup \beta$ (non-deterministic choice) and $\alpha;\beta$ (sequential combination) are actions;*
3. *If $\alpha$ is an action, then $\alpha^*$ (iterate $\alpha$ zero or more times) is an action as well.*

`Definition` **2.1.2** *The set of* formulas *is the smallest set produced by the following rules:*

1. *For a formula $\eta$ of the agent's internal language, $\mathbf{B}\eta$ is a formula of the external language;*

2. *If $\phi$ and $\psi$ are formulas, then $\neg\phi$ and $\phi \vee \psi$ are formulas;*

3. *If $\phi$ is a formula and $\alpha$ is an action, then $\langle\alpha\rangle\phi$ is a formula.*

The connectives $\wedge$, $\rightarrow$, $\leftrightarrow$ and the modality symbol $[\alpha]$ are defined in the usual way.

Now the semantics.

`Definition` **2.1.3** *A* state *is a finite set of formulas in the agent's internal language.*

`Definition` **2.1.4** *Accessibility relations between states induced by the actions, $[\![\alpha]\!]$, are built recursively:*

1. *For an elementary action (that is, a rule)* R*, states $s$ and $t$ stand in the relation $[\![\mathrm{R}]\!]$ iff either*

   (a) $t = s$,[3] *or*

   (b) $t = s \cup \{\eta\}$*, and either*

      i. R *is a simple rule, and there exists an instance of* R *such that its conclusion is $\eta$ and all the premises belong to $s$, or*

[3] The reason I include the first alternative (which makes every accessibility relation in the system reflexive) is that my system is meant to model agents such as humans. It is safe to assume that beyond a certain class of beliefs that we are interested in, the agent also possesses some unrelated beliefs which he can also reason about. Such unrelated inferences will not change the parts of his state that we are set to describe. In addition, the agent is free to perform inferences whose conclusions are already in his belief state.

*ii.* R *is a subproof rule, and there exists an instance of* R *such that its conclusion is* $\eta$*, and all the formulas used in the subproof as premises, except for the opening one, belong to* $s$*.*

2. *Accessibility relation induced by a non-deterministic choice action is the union of relations induced by its components:*

$$[\![\alpha \cup \beta]\!] = [\![\alpha]\!] \cup [\![\beta]\!]$$

3. *Relation induced by a sequential action is the composition of relations induced by its components:*

$$[\![\alpha; \beta]\!] = [\![\alpha]\!] \circ [\![\beta]\!]$$

4. *Accessibility relation induced by an iterated action is the reflexive transitive closure of that action's relation:*

$$[\![\alpha^*]\!] = [\![\alpha]\!]^* = \bigcup_{i \in \mathit{0..\omega}} [\![\alpha]\!]^i$$

`Definition` **2.1.5** *Interpretation function for formulas relative to a state,* $[\![\phi]\!]^s$*, is defined by the rules:*

1. *An elementary formula* $\mathbf{B}\eta$ *is true at* $s$ *iff* $\eta$ *belongs to* $s$*:*

$$[\![\mathbf{B}\eta]\!]^s = \begin{cases} \textbf{true} & \textit{if } \eta \in s \\ \textbf{false} & \textit{otherwise} \end{cases}$$

2. *Truth functional connectives are defined as usual:*

$$\begin{array}{lcl} [\![\neg\phi]\!]^s & = & \neg[\![\phi]\!]^s \\ [\![\phi \vee \psi]\!]^s & = & [\![\phi]\!]^s \vee [\![\psi]\!]^s \end{array}$$

*3. Diamonds are also defined as usual, with accessibility relations corresponding to the actions:*

$$[\![\langle\alpha\rangle\phi]\!]^s = \begin{cases} \textbf{true} & \textit{if there exists a state } t \textit{ such that} \\ & s[\![\alpha]\!]t \textit{ and } [\![\phi]\!]^t = \textbf{\textit{true}} \\ \textbf{false} & \textit{otherwise} \end{cases}$$

Essentially, my logic has a single Kripke-style model for dynamic logic $M = \langle S, R_0, \ldots, R_n, V\rangle$ where $S$ is a (countably infinite) set of states, $R_0$ to $R_n$ are accessibility relations for elementary actions, and $V$ is a labeling function assigning sets of elementary formulas to states. I could have stated explicitly that a model is such a structure and impose restrictions on models that guarantee, for each state $s$ and each rule instance $\eta_1, \ldots \eta_n \Rightarrow_{\mathrm{R}_i} \zeta$ applicable at that state, that there is a state $t$ accessible from $s$ through $R_i$ such that $V(t) = V(s) \cup \{\zeta\}$. Furthermore, for each arc in the accessibility relation $sR_it$ there should be a corresponding rule instance. Similar restrictions should be given for subproof rules. This is the approach taken in (Jago, 2006).

But for such a setup, it is easy to show that states with identical labelings are bisimilar, and thus modally equivalent, so essentially all the information you need in a state is the set of formulas believed there. I chose to identify a state with that set of formulas.

#### 2.1.0.1 Examples

To provide a specific example, let us assume that the agent's internal language $S$ is the standard language of sentential logic (with $\neg$, $\wedge$, $\vee$ and $\rightarrow$ as connectives), and that the agent uses the following variant of natural deduction as his inference system:

$$\begin{array}{c}\eta\\ \ldots\\ \theta\wedge\neg\theta\\ \hline \neg\eta\end{array}\ \neg\mathbf{I} \qquad\qquad \frac{\neg\neg\eta}{\eta}\ \neg\mathbf{E}$$

$$\frac{\eta \qquad \zeta}{\eta\wedge\zeta}\ \wedge\mathbf{I} \qquad\qquad \frac{\eta\wedge\zeta}{\eta}\ \wedge\mathbf{E} \quad \frac{\eta\wedge\zeta}{\zeta}\ \wedge\mathbf{E}$$

$$\frac{\eta}{\eta\vee\zeta}\ \vee\mathbf{I} \quad \frac{\zeta}{\eta\vee\zeta}\ \vee\mathbf{I} \quad \frac{\eta\vee\zeta \qquad \eta\rightarrow\theta \qquad \zeta\rightarrow\theta}{\theta}\ \vee\mathbf{E}$$

$$\begin{array}{c}\eta\\ \ldots\\ \zeta\\ \hline \eta\rightarrow\zeta\end{array}\ \rightarrow\mathbf{I} \qquad\qquad \frac{\eta\rightarrow\zeta \qquad \eta}{\zeta}\ \rightarrow\mathbf{E}$$

$$\frac{}{\eta}\ \mathbf{O}$$

(The **O** is for “observation”.)

Let us also introduce the following abbreviation:

$$\text{Infer} \equiv (\neg\mathbf{I}\cup\neg\mathbf{E}\cup\wedge\mathbf{I}\cup\wedge\mathbf{E}\cup\vee\mathbf{I}\cup\vee\mathbf{E}\cup\rightarrow\mathbf{I}\cup\rightarrow\mathbf{E})^*$$

(This is the transitive closure of all the usual rules, not including **O**.)

We also introduce a new symbol. Let $!\phi$ mean “$\phi$ has become true at the last state change”. It is not quite an abbreviation — capturing this idea would require extending our current semantics to keep track of the formula that has just been introduced, — but we can always recursively translate a formula of the form $\langle \mathrm{R}\rangle\phi$ containing the ! symbol into one that does not:

$$\begin{array}{lcl}
\langle \mathrm{R}\rangle !\phi & \equiv & \neg\phi \wedge \langle \mathrm{R}\rangle\phi \\
\langle \mathrm{R}\rangle(!\phi \vee \psi) & \equiv & \langle \mathrm{R}\rangle !\phi \vee \langle \mathrm{R}\rangle\psi \\
\langle \mathrm{R}\rangle(\phi \vee !\psi) & \equiv & \langle \mathrm{R}\rangle\phi \vee \langle \mathrm{R}\rangle !\psi \\
\langle \mathrm{R}\rangle\neg !\phi & \equiv & \phi \vee \langle \mathrm{R}\rangle\neg\phi \\
\langle \mathrm{R}\rangle\langle \mathrm{Q}\rangle !\phi & \equiv & \langle \mathrm{R}\rangle(\langle Q\rangle !\phi) \\
\langle \mathrm{R}\rangle\langle \alpha \cup \beta\rangle !\phi & \equiv & \langle \mathrm{R}\rangle\langle \alpha\rangle !\phi \vee \langle \mathrm{R}\rangle\langle \beta\rangle !\phi \\
\langle \mathrm{R}\rangle\langle \alpha ; \beta\rangle !\phi & \equiv & \langle \mathrm{R}\rangle\langle \alpha\rangle\langle \beta\rangle !\phi \\
\langle \mathrm{R}\rangle\langle \alpha^*\rangle !\phi & \equiv & \langle \mathrm{R}\rangle !\phi \vee \langle \mathrm{R}\rangle\langle \alpha^*\rangle\langle \alpha\rangle !\phi \\
\langle \mathrm{R}\rangle !!\phi & \equiv & \langle \mathrm{R}\rangle !\phi \\
\langle \mathrm{R}\rangle\neg !!\phi & \equiv & \langle \mathrm{R}\rangle\neg !\phi
\end{array}$$

Here are some formulas in our language, with their intended meaning:

| | |
|---|---|
| $\langle \to\mathbf{E}\rangle !\mathbf{B}p$ | $p$ is derivable by a single application of $\to\mathbf{E}$ |
| $\langle \mathrm{Infer}\rangle\mathbf{B}p$ | $p$ is derivable (it is not specified what the derivation looks like). |
| $\langle \wedge\mathbf{E}; \vee\mathbf{E}; \to\mathbf{E}\rangle !\mathbf{B}p$ | There is a derivation with a specific shape that leads to $p$. |
| $[\mathrm{Infer}](\mathbf{B}p \to \mathbf{B}q)$ | The only way the agent can reach $p$ is via $q$. |
| $[\mathbf{O}](\mathbf{B}q \to !\langle \mathrm{Infer}\rangle\mathbf{B}p)$ | If the agent acquires the belief $q$, he will gain the ability to derive $p$. |

Here are some valid formula schemas:[4]

$$[\alpha](\phi \to \psi) \to ([\alpha]\phi \to [\alpha]\psi) \tag{2.1}$$

$$\langle \alpha\rangle(\phi \vee \psi) \leftrightarrow \langle \alpha\rangle\phi \vee \langle \alpha\rangle\psi \tag{2.2}$$

$$\langle \alpha^*\rangle\phi \leftrightarrow \phi \vee \langle \alpha\rangle\langle \alpha^*\rangle\phi \tag{2.3}$$

$$\mathbf{B}\eta \to [\mathrm{Infer}]\mathbf{B}\eta \tag{2.4}$$

$$\langle \mathrm{R}\rangle(!\mathbf{B}\eta \to \neg !\mathbf{B}\zeta), \quad \text{for } \eta \not\equiv \zeta \tag{2.5}$$

$$\mathbf{B}\neg\neg\eta \to \langle \neg\mathbf{E}\rangle\mathbf{B}\eta \tag{2.6}$$

[4] $\bot$ is a shortcut for an arbitrary contradiction.

$$\langle \neg\mathbf{E}\rangle !\mathbf{B}\eta \rightarrow \mathbf{B}\neg\neg\eta \tag{2.7}$$

$$[\mathbf{O}](\mathbf{B}\eta \rightarrow \langle \text{Infer}\rangle \mathbf{B}\bot) \rightarrow \langle \neg\mathbf{I}\rangle \mathbf{B}\neg\eta \tag{2.8}$$

$$\langle \neg\mathbf{I}\rangle !\mathbf{B}\neg\eta \rightarrow [\mathbf{O}](\mathbf{B}\eta \rightarrow \langle \text{Infer}\rangle \mathbf{B}\bot) \tag{2.9}$$

$$\mathbf{B}\eta \wedge \mathbf{B}\zeta \rightarrow \langle \wedge\mathbf{I}\rangle \mathbf{B}(\eta \wedge \zeta) \tag{2.10}$$

$$\langle \wedge\mathbf{I}\rangle !\mathbf{B}(\eta \wedge \zeta) \rightarrow \mathbf{B}\eta \wedge \mathbf{B}\zeta \tag{2.11}$$

$$\mathbf{B}(\eta \wedge \zeta) \rightarrow \langle \wedge\mathbf{E}\rangle \mathbf{B}\eta \wedge \langle \wedge\mathbf{E}\rangle \mathbf{B}\zeta \tag{2.12}$$

$$!\langle \wedge\mathbf{E}\rangle \mathbf{B}\eta {\wedge} !\langle \wedge\mathbf{E}\rangle \mathbf{B}\zeta \rightarrow !\mathbf{B}(\eta \wedge \zeta) {\vee} !\mathbf{B}(\zeta \wedge \eta) \tag{2.13}$$

$$\mathbf{B}\eta \rightarrow \langle \vee\mathbf{I}\rangle \mathbf{B}(\eta \vee \zeta) \wedge \langle \vee\mathbf{I}\rangle \mathbf{B}(\zeta \vee \eta) \tag{2.14}$$

$$\langle \vee\mathbf{I}\rangle !\mathbf{B}(\eta \vee \zeta) \rightarrow \mathbf{B}\eta \vee \mathbf{B}\zeta \tag{2.15}$$

$$\mathbf{B}(\eta \vee \zeta) \wedge \mathbf{B}(\eta \rightarrow \theta) \wedge \mathbf{B}(\zeta \rightarrow \theta) \rightarrow \langle \vee\mathbf{E}\rangle \mathbf{B}\theta \tag{2.16}$$

$$[\mathbf{O}](\mathbf{B}\eta \rightarrow \langle \text{Infer}\rangle \mathbf{B}\zeta) \rightarrow \langle {\rightarrow}\mathbf{I}\rangle \mathbf{B}(\eta \rightarrow \zeta) \tag{2.17}$$

$$\langle tI\rangle !\mathbf{B}(\eta \rightarrow \zeta) \rightarrow [\mathbf{O}](\mathbf{B}\eta \rightarrow \langle \text{Infer}\rangle \mathbf{B}\zeta) \tag{2.18}$$

$$\mathbf{B}(\eta \rightarrow \zeta) \wedge \mathbf{B}\eta \rightarrow \langle {\rightarrow}\mathbf{E}\rangle \mathbf{B}\zeta \tag{2.19}$$

$$\langle \mathbf{O}\rangle \mathbf{B}\eta \tag{2.20}$$

(2.1) is the familiar axiom **K**; putting it here reminds that our logic is a normal modal logic; (2.2) and (2.3) are examples of propositional dynamic logic axioms; all other PDL axioms, of course, are valid as well. (2.4) states that the agent's reasoning is monotonic; beliefs are never lost. (2.5) declares that the agent acquires at most one belief at each inferential step. The rest of the formula schemas reflect the behaviour of natural deduction rules. (2.6), (2.8), (2.10), (2.12), (2.14), (2.16), (2.17), (2.19) and (2.20) demonstrate "forward reasoning": given that the agent knows the premises of a rule instance, we conclude that he is able to deduce the conclusion. We can state such axioms for every rule. On the other hand, (2.7), (2.9), (2.11), (2.13), (2.15) and (2.18) show "backward reasoning": knowing that by application of a given rule the agent can deduce a certain new formula (hence the !s in all of these rules), we conclude that the agent already possesses the premises.[5] For certain rules there is no way to recover the premises from the conclusion ($\vee\mathbf{E}$, $\rightarrow\mathbf{I}$, $\mathbf{O}$). For example, given that $\eta$ is derivable by an application of *modus ponens*, we can conclude that for *some* $\zeta$, the agent has both $\zeta$ and $\zeta \rightarrow \eta$ in his belief box, but we don't know what $\zeta$ is. Since we don't have quantification over formulas in our language, we cannot even express that knowledge. This consideration makes it unlikely for $F_s$ to turn out axiomatizable.

Monotonicity of the agent's inference system ensures that certain formulas, such as those of the form $\langle\alpha\rangle\mathbf{B}\eta$, once true stay true. That is, on any chain of states linked by transitions, these formulas can change their value at most once — from **false** to **true**. Of course, there is a dual class, formulas of the form $[\alpha]\neg\mathbf{B}\eta$, that only change their value from **true** to **false**.

A question emerges: is it possible for a formula to change its truth value an infinite number of times as the agent's inference progresses? The answer is yes. For example, the formula $F = \langle\rightarrow\mathbf{E}\rangle(\langle\rightarrow\mathbf{E}\rangle\mathbf{B}p \wedge \neg\langle\rightarrow\mathbf{E}\rangle\mathbf{B}q)$ does that in the following chain of inference, starting from the state $s_0 = \{s, s \rightarrow r, r \rightarrow$

[5] In (Konolige, 1990) this mode of reasoning is called *explanatory* belief ascription.

$p, r \to q\}$:

| Rule used | Result | Value of $F$ |
|---|---|---|
| At $s_0$ | | **false** |
| $\to$**I** | $(r \wedge r) \to p$ | **false** |
| $\to$**I** | $s \to (r \wedge r)$ | **true** |
| $\to$**I** | $(r \wedge r) \to q$ | **false** |
| $\to$**I** | $(r \wedge (r \wedge r)) \to p$ | **false** |
| $\to$**I** | $s \to (r \wedge (r \wedge r))$ | **true** |
| $\to$**I** | $(r \wedge (r \wedge r)) \to q$ | **false** |
| $\ldots$ | | |

Existence of such formulas makes it impossible to attack the problem of decidability for $F_S$ in the following way: create a finite submodel by filtering the whole model through some set of sentences, then enrich it so that every nontrivial transition is justified by some instance of the corresponding rule, then add other nodes so that the submodel becomes part of the whole model $M$. When the set of sentences we filter through includes formulas such as $F$, the filtered structure is going to contain nontrivial cycles. Therefore, no 'enrichment' will ever convert it into a part of a proper model.

For certain applications, we might want to ascribe our agents ability to perform certain simple inferences "effortlessly", so that an agent who possesses the premises of such an inference automatically acquires the conclusion. (For example, our agent may perform conjunction simplifications in such a way.) To capture such "free" inferences, we might redefine the notion of a state: a state is a set of sentences closed under certain operations. This, in turn, leads to a modified definition of the accessibility relation for elementary actions:

For a rule R, states $s$ and $t$ stand in the relation $[\![\mathrm{R}]\!]$ iff either

1. $t = s$, or

2. R is a simple rule, there exists an instance of R where all the premises belong to $s$, and $t$ is the smallest state extending $s$ and containing the conclusion $\eta$ of that instance.
3. R is a subproof rule, there exists an instance of R such that all the premises mentioned in the subproof belong to $s$, and $t$ is the smallest state extending $s$ and containing the conclusion $\eta$ of that instance.

This move is proposed by (Ågotnes, 2004). I will not use it in the remainder of this text.

The logic as described here lacks compactness for two separate reasons. On the one hand, it is a species of dynamic logic, and so the following kind of finitely satisfiable but contradictory sets of sentences can be built:

$$\{\langle a^*\rangle \mathbf{B}p, \neg\mathbf{B}p, [a]\neg\mathbf{B}p, [a^2]\neg\mathbf{B}p, [a^3]\neg\mathbf{B}p\ldots\}$$

The reason for lack of compactness here is the presence of the $^*$ operator, which is equivalent to a disjunction of countably many formulas. On the other hand, the fact that our intended model deals with a countable number of sentences in the agent's language, and all of them are nameable, leads to another way of building an inconsistent set whose each finite subset is consistent. Let $\rightarrow\!\mathbf{E}$ be the name of the *modus ponens* rule of inference. Then the set

$$S = \{\neg\mathbf{B}p, \langle\rightarrow\!\mathbf{E}\rangle\mathbf{B}p\} \cup \{\neg\mathbf{B}(\eta \rightarrow p) \mid \eta \in \mathcal{L}\}$$

is inconsistent. Ågotnes (2004) deals with this particular kind of compactness failure by stipulating that the agent's language contains a formula that is not nameable in the external language. In that case, the set $S$ becomes consistent. But since in my case the other kind of compactness failure would still remain, I don't see any reason to adopt this measure for my system.

## 2.2 Logic for a Gentzen-style reasoning agent

One reason to be dissatisfied with the system of Section 2.1 is that there is no way to restrict the content of subproofs. For example, given $F_S$, the action $\neg\mathbf{I}; \neg\mathbf{E}$ (indirect proof followed by double-negation elimination) is equivalent to an arbitrary derivation Infer. In order to restrict the content of subproofs, we would need to make the action that describes the embedded derivation a parameter of the subproof action. But in that case, the iteration construct turns out to be insufficiently expressive.

Besides that, we would like to use our system to describe the reasoning of human agents. One obstacle to doing that is that positing a particular derivation system as the reasoning mechanism humans actually use seems unnecessarily ambitious. We would like the logic to be flexible enough, so that it can model various derivation systems. In principle, this seems doable. A step in the proof according to one system may be reproduced as a number of steps in another. So, a typical statement like

$$\langle \mathrm{R}^* \rangle \mathbf{B}\eta$$

which means "$\eta$ is derivable by repeated use of the rule R", would still be expressible even if the agent's system does not have R, but can model it as a sequence of rules $\mathrm{Q}; \mathrm{S}; \mathrm{T}$:[6]

$$\langle (\mathrm{Q}; \mathrm{S}; \mathrm{T})^* \rangle \mathbf{B}\eta$$

Unfortunately, in the logic of Section 2.1 this strategy will not work: its expressive power is less than it seems. It turns out that the internal structure

[6] This imitation would still be incomplete, because intermediate results of applying Q and $\mathrm{Q}; \mathrm{S}$ would accumulate in the agent's state.

of an iterated action plays no role; all that matters is what rules are mentioned in it.

Lemma **2.2.1** *For an action $\alpha$, let $mash(\alpha)$ be the disjunction of all elementary actions (i. e., rules) mentioned in $\alpha$. Then for any formula $\phi$, $\langle\alpha^*\rangle\phi$ is equivalent to $\langle mash(\alpha)^*\rangle\phi$.*

Proof: Suppose $\langle\alpha^*\rangle\phi$ is true at state $s$. Then there is some state $t$ such that $\phi$ is true at $t$, and there is a finite sequence of states $s = s_0, s_1, \ldots s_n = t$ such that each $s_i$ is connected to $s_{i+1}$ by $\alpha$. Each of these $\alpha$ connections, in turn, is decomposable into a finite sequence of elementary action connections by rules mentioned in the definition of $\alpha$. Therefore, there is a finite sequence of states connected by elementary actions in $mash(\alpha)$ leading from $s$ to $t$. But this is precisely what it means for $s$ and $t$ to be connected by $mash(\alpha)^*$. Therefore, $\langle mash(\alpha)\rangle\phi$ is true at $s$.

On the other hand, suppose $\langle mash(\alpha)^*\rangle\phi$ is true at $s$. This means there is a sequence of states connected by $mash(\alpha)$ and leading to some $t$, with $\phi$ true at $t$. Now we shall prove that each $mash(\alpha)$ link is also an $\alpha$ link. Every such link is a rule action R from some state $s_i$ to $s_{i+1}$. We build the proof by induction on the complexity of $\alpha$.

- If $\alpha$ is a rule Q, then $mash(\mathrm{Q}) = \mathrm{Q}$.
- If $\alpha$ has the form $\beta \cup \gamma$, then either $\beta$ or $\gamma$ contain R, and so, by the inductive hypothesis, $s_i[\![\beta]\!]s_{i+1}$ or $s_i[\![\gamma]\!]s_{i+1}$.
- If $\alpha$ has the form $\beta;\gamma$, then, again, either $\beta$ or $\gamma$ contain R. Assume it's $\beta$. Then $s_i[\![\beta]\!]s_{i+1}$ by the inductive hypothesis. But since all our action

relations are reflexive, $s_{i+1}[\![\gamma]\!]s_{i+1}$, therefore, $s_i[\![\beta;\gamma]\!]s_{i+1}$. Similarly for the case where $\gamma$ contains R.

- If $\alpha$ has the form $\beta^*$, then, by the inductive hypothesis, $s_i[\![\beta]\!]s_{i+1}$, and therefore, by definition, $s_i[\![\beta^*]\!]s_{i+1}$. □

These considerations lead us to search for a system that would be able to reflect the structure of derivations in greater detail. One way to do this is to switch the agent's inference system from Fitch-style natural deduction to Gentzen-style. As a result, we will be able to build our actions as skeletons of derivation trees. This, however, forces us to make both syntax and semantics of our logic more complicated.

In our new logic $G_{\mathcal{L}}$ we assume that the rules of inference have the following form:

$$\frac{\Gamma, \zeta_1 \vdash \eta_1 \qquad \dots \qquad \Gamma, \zeta_k \vdash \eta_k \qquad \Gamma \vdash \eta_{k+1} \qquad \dots \qquad \Gamma \vdash \eta_m}{\Gamma \vdash \theta}\ \mathrm{R}$$

That is, whenever the conclusion is taken to follow from a set of sentences $\Gamma$, some premises ($\eta_{k+1}$ to $\eta_m$) need to follow from the same set, and some ($\eta_1$ to $\eta_k$) follow from the same set extended by one additional premise (different for each subproof).

Second, we have a new notion of what an action is and how actions are combined:

**Definition 2.2.2** *Let $x, y, \dots$ be a set of* action variables*, distinct from every other syntactic object we used before.*

1. *An action variable is an action;*

*2. $\epsilon$ (the empty action) is an action;*

*3. $A$ is an action (to be discussed shortly);*

*4. If $\alpha_1, \alpha_2, \ldots \alpha_m$ are actions, and* R *is a rule with $m$ premises, then*

$$\frac{\alpha_1 \qquad \alpha_2 \ldots \qquad \alpha_m}{} \mathrm{R}$$

*is an action;*

*5. If $\alpha$ and $\beta$ are actions, then $\alpha \cup \beta$ and $\alpha; \beta$ are actions;*

*6. If $\alpha$ is an action and $x$ is an action variable, then $\mu x.\alpha$ is an action. Variable $x$ is* bound *in that action.*

We only allow *closed* action terms as action prefixes in formulas — where all the action variables are bound, according to the usual definition.

The expression $\mu x.\alpha$ denotes an action that is defined recursively — that is, an action $\beta$ which is identical to $\alpha$, except that all occurrences of $x$ within $\alpha$ are replaced by $\beta$ itself. This bit of formalism is taken from CFDL (context-free dynamic logic, (Harel, 1979)); I use it in order to control the patterns in derivation trees more finely. The old iteration construct can be defined as

$$\alpha^* \equiv \mu x.(\epsilon \cup (x; \alpha))$$

We also add a new elementary formula type, $A\eta$, meaning 'the agent's attention is directed at $\eta$'.

We now need to specify the semantics for our modified logic. First of all, the notion of state is modified.

`Definition` **2.2.3** *A* state *is a pair $s = \langle \Gamma, \eta \rangle$, where $\Gamma$ is a finite set of sentences in the agent's internal language, and $\eta \in \Gamma$ is a sentence ('the focus of attention'). We shall denote $\Gamma$ as $\Gamma(s)$ and $\eta$ as $A(s)$.*

One can also add the empty set as an exceptional state where no focus of attention exists.

Formula $A\eta$ is true at a state $s$ whenever $A(s) = \eta$.

For a state $s = \langle \Gamma, \eta \rangle$ and a sentence $\zeta$, we shall denote the state $\langle \Gamma \cup \{\zeta\}, \zeta \rangle$ as $s + \zeta$.

Now we redefine the accessibility relation between states.

`Definition` **2.2.4** *1. For an empty action, $s \llbracket \epsilon \rrbracket t$ iff $s = t$;*

*2. For the action $A$ (refocus attention), for any $\Gamma$ and any $\eta_1, \eta_2 \in \Gamma$, $\langle \Gamma, \eta_1 \rangle \llbracket A \rrbracket \langle \Gamma, \eta_2 \rangle$;*

*3. For a rule* R *with $m$ premises, actions $\alpha_1, \ldots \alpha_m$, and states $s$ and $t$,*

$$s \llbracket \frac{\alpha_1 \ldots \qquad \alpha_m}{} \ \mathrm{R} \rrbracket t$$

*iff there is an instance of* R *of the form*

$$\frac{\Gamma, \sigma\zeta_1 \vdash \sigma\eta_1 \ldots \qquad \Gamma, \sigma\zeta_k \vdash \sigma\eta_k \qquad \Gamma \vdash \sigma\eta_{k+1} \ldots \qquad \Gamma \vdash \sigma\eta_m}{\Gamma \vdash \sigma\theta} \ \mathrm{R},$$

*such that for each $i \in 1 \ldots k$ there is some $t_i$ such that $(s + \sigma\zeta_i) \llbracket \alpha_i \rrbracket t_i$ and $A(t_i) = \sigma\eta_i$, for each $j \in (k+1) \ldots m$ there is some $t_j$ such that $s \llbracket \alpha_j \rrbracket t_j$ and $A(t_j) = \sigma\eta_j$, and $t = \langle \Gamma(s) \cup \bigcup_{j \in (k+1 \ldots m)} \Gamma(t_j) \cup \{\sigma\theta\}, \sigma\theta \rangle$.*[7]

[7] We accumulate intermediate results of subproofs without additional premises in the final state of the proof.

4. *Accessibility relation induced by a non-deterministic choice action is the union of relations induced by its components:*

$$[\![\alpha \cup \beta]\!] = [\![\alpha]\!] \cup [\![\beta]\!]$$

5. *Accessibility relation induced by a sequential action is the composition of relations induced by its components:*

$$[\![\alpha ; \beta]\!] = [\![\alpha]\!] \circ [\![\beta]\!]$$

6. *Accessibility relation for a recursive action $\mu x.\alpha$ is computed in the following way.*

   *For a relation $R$, let $[\![\alpha_{x/R}]\!]$ be the accessibility relation for $\alpha$ where $x$ is taken to have $R$ as its accessibility relation.*

   *Let us build a sequence of relations $R_i$ with $R_0 = \emptyset$ and*

$$R_{i+1} = R_i \cup [\![\alpha_{x/R_i}]\!]$$

   *Now,*

$$[\![\mu x.\alpha]\!] = \bigcup_i R_i$$

#### 2.2.0.2 Examples

An agent whose internal language is sentential logic could use the following set of natural deduction rules:

$$\frac{\Gamma, \eta \vdash \bot}{\Gamma \vdash \neg\eta}\ \neg\mathbf{I} \qquad\qquad \frac{\Gamma \vdash \neg\neg\eta}{\Gamma \vdash \eta}\ \neg\mathbf{E}$$

$$\frac{\Gamma \vdash \eta \qquad \Gamma \vdash \zeta}{\Gamma \vdash \eta \wedge \zeta}\ \wedge\mathbf{I} \qquad\qquad \frac{\Gamma \vdash \eta \wedge \zeta}{\Gamma \vdash \eta}\ \wedge\mathbf{E} \quad \frac{\Gamma \vdash \eta \wedge \zeta}{\Gamma \vdash \zeta}\ \wedge\mathbf{E}$$

$$\frac{\Gamma \vdash \eta}{\Gamma \vdash \eta \vee \zeta}\ \vee\mathbf{I} \quad \frac{\Gamma \vdash \zeta}{\Gamma \vdash \eta \vee \zeta}\ \vee\mathbf{I} \quad \frac{\Gamma \vdash \eta \vee \zeta \qquad \Gamma, \eta \vdash \theta \qquad \Gamma, \zeta \vdash \theta}{\Gamma \vdash \theta}\ \vee\mathbf{E}$$

$$\frac{\Gamma, \eta \vdash \zeta}{\Gamma \vdash \eta \to \zeta}\ {\to}\mathbf{I} \qquad\qquad \frac{\Gamma \vdash \eta \to \zeta \qquad \Gamma \vdash \eta}{\Gamma \vdash \zeta}\ {\to}\mathbf{E}$$

$$\frac{}{\eta}\ \mathbf{O}$$

Our new language $G_S$ is strictly more expressive than $F_S$. Here is a formula which cannot be translated to $F_S$:

$$\langle \mu x.(A \cup \frac{A \qquad x}{\phantom{A}}\ {\to}\mathbf{E}))\rangle \mathbf{B}p$$

It means that the agent can derive $p$ by using a series of *modus ponens* steps, and the result of the previous step always plays the role of the second premise in the next step. This proof, for example, meets the description:

$$\frac{q \to p \qquad \dfrac{r \to q \qquad \dfrac{s \to r \qquad r}{r}\ {\to}\mathbf{E}}{q}\ {\to}\mathbf{E}}{p}\ {\to}\mathbf{E}$$

And this one does not:

$$\dfrac{q \to p \qquad \dfrac{\dfrac{s \to (r \to q) \qquad s}{r \to q}\,{\to}\mathbf{E} \qquad r}{q}\,{\to}\mathbf{E}}{p}\,{\to}\mathbf{E}$$

The rules of $F_s$ can be trivially reproduced in $G_s$:

$$\neg\mathbf{I}_{F_S} = \dfrac{\text{Infer}}{}\ \neg\mathbf{I} \qquad\qquad \neg\mathbf{E}_{F_S} = \dfrac{A}{}\ \neg\mathbf{E}$$

$$\wedge\mathbf{I}_{F_S} = \dfrac{A \qquad A}{}\ \wedge\mathbf{I} \qquad\qquad \wedge\mathbf{E}_{F_S} = \dfrac{A}{}\ \wedge\mathbf{E}$$

$$\vee\mathbf{I}_{F_S} = \dfrac{A}{}\ \vee\mathbf{I} \qquad \vee\mathbf{E}_{F_S} = \dfrac{A \qquad \dfrac{A \qquad A}{}\ {\to}\mathbf{E} \qquad \dfrac{A \qquad A}{}\ {\to}\mathbf{E}}{}\ \vee\mathbf{E}$$

$${\to}\mathbf{I}_{F_S} = \dfrac{\text{Infer}}{}\ {\to}\mathbf{I} \qquad\qquad {\to}\mathbf{E}_{F_S} = \dfrac{A \qquad A}{}\ {\to}\mathbf{E}$$

where

$$\begin{aligned}\text{Infer} = \mu x.\quad & (A \cup \dfrac{x}{}\ \neg\mathbf{I} \cup \dfrac{x}{}\ \neg\mathbf{E} \cup \dfrac{x \qquad x}{}\ \wedge\mathbf{I} \cup \dfrac{x}{}\wedge\mathbf{E}\cup \\ & \dfrac{x}{}\ \vee\mathbf{I} \cup \dfrac{x \qquad x \qquad x}{}\ \vee\mathbf{E} \cup \dfrac{x}{}\ {\to}\mathbf{I} \cup \dfrac{x \qquad x}{}\ {\to}\mathbf{E})\end{aligned}$$

### 2.2.1 Undecidability

Now that, using the attention mechanism, we can demand in our action specifications that the next step in the derivation be applied to the result of the previous step, it's easy to prove that our current system is, at least in general, undecidable.

We do this by converting our actions into programs for register machines (Lambek, 1961; Boolos et al., 2002) – a variation of Turing machines. The halting problem for register machines is undecidable.

A register machine state is composed of a memory and a program. Memory contains a finite number of cells (registers), each capable of holding a natural number. The program is a finite sequence of instructions. There are three types of instructions:

- $+_i\ n$ — add 1 to the contents of register $i$, perform instruction number $n$ next;
- $-_i\ n, m$ — if the number contained in register $i$ is nonzero, decrease the contents by 1 and perform instruction number $n$; otherwise perform instruction number $m$;
- **halt** — stop execution.

The machine starts at instruction number 0.

We can now specify translation of register machine states into formulas of our dynamic logic. Formulas of the agent's internal language will correspond to memory states of the register machine under consideration. Namely, each memory state $S = \langle x_1, \ldots, x_n \rangle$ will be represented by a formula

$$T(S) = x_1 \bullet \ldots \bullet x_n$$

where each $x_i$ is represented in the standard unary notation – a sequence of $x_i$ $s$ symbols ending in 0. Any agent's state $\langle \Gamma, T(S) \rangle$ represents the machine state $S$ (that is, we are only interested in the formula under attention; other beliefs play no role in our representation). A register machine program will be encoded as an action of our dynamic logic.

Elementary rules used in the translation will be the following:

$$\frac{a_1 \bullet \dots \bullet a_{j-1} \bullet a_j \bullet a_{j+1} \bullet \dots a_n}{a_1 \bullet \dots \bullet a_{j-1} \bullet sa_j \bullet a_{j+1} \bullet \dots a_n}\ \textsc{a}_{+,j}$$

$$\frac{a_1 \bullet \dots \bullet a_{j-1} \bullet sa_j \bullet a_{j+1} \bullet \dots a_n}{a_1 \bullet \dots \bullet a_{j-1} \bullet a_j \bullet a_{j+1} \bullet \dots a_n}\ \textsc{a}_{-,j}$$

$$\frac{a_1 \bullet \dots \bullet a_{j-1} \bullet 0 \bullet a_{j+1} \bullet \dots a_n}{a_1 \bullet \dots \bullet a_{j-1} \bullet 0 \bullet a_{j+1} \bullet \dots a_n}\ \textsc{a}_{0,j}$$

Rule $\textsc{a}_{+,j}$ represents adding 1 to the $j$-th register; $\textsc{a}_{-,j}$ represents subtracting 1 from it, and is only applicable to a state where the content of the register is nonzero; $\textsc{a}_{0,j}$ does nothing, and it's only applicable when the content of the $j$-th register is 0.

Since these are rules, their names become actions only when given another action to produce a premise. Let us write $a_{+,j}$ for $\overset{\epsilon}{-}\ \textsc{a}_{+,j}$, $a_{-,j}$ for $\overset{\epsilon}{-}\ \textsc{a}_{-,j}$, and $a_{0,j}$ for $\overset{\epsilon}{-}\ \textsc{a}_{0,j}$, — that is, for actions where the rules are applied to the current focus of attention.

Translation is defined with the help of the following procedure $T_M(j, c)$, where $j$ is the node's index and $c$, a context, is a finite set of numbers indicating variables that are taken to be already defined. We write $cj$ for a context extended with $j$ – that is, for $c \cup \{j\}$.

- If $j \in c$, then $T_M(j, c) = x_j$. Otherwise,
- if the $j$-th instruction is **halt**, then $T_M(j, c) = \epsilon$;
- if the $j$-th instruction is $+_i\ n$, then

$$T_M(j, c) = \mu x_j.(a_{+,j}; T_M(n, cj))$$

- if the $j$-th instruction is $-_i\, n, m$, then

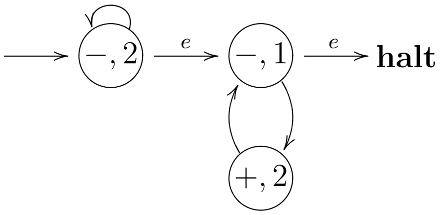
$$T_M(j, c) = \mu x_j.((a_{-,j}; T_M(n, cj)) \cup (a_{0,j}; T_M(m, cj)))$$

Translation for the whole program is defined as $T_M = T_M(0, \emptyset)$

For example, consider the following machine

$$\begin{array}{rl} M = \langle & (-_2 0, 1), \\ & (-_1 2, 3), \\ & (+_2 1), \\ & \mathbf{halt} \quad \rangle \end{array}$$

whose behaviour can be described as "move the contents of register 1 into register 2". Here is a diagram showing its state transitions:



Our translation procedure gives the following result for $M$:

$$T_M = \mu x_0.((a_{-,2}; x_0) \cup (a_{0,2}; \mu x_1.((a_{-,1}; \mu x_2.(a_{+,2}; x_1)) \cup (a_{0,1}; \epsilon))))$$

which is equivalent, after dropping unused variables and the empty action, to

$$\mu x_0.((a_{-,2}; x_0) \cup (a_{0,2}; \mu x_1.((a_{-,1}; a_{+,2}; x_1) \cup a_{0,1})))$$

Starting with a state $s_0 = \langle \{ss0 \bullet s0\}, ss0 \bullet s0 \rangle$, the following derivation corresponds to the run of the machine:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{ss0 \bullet s0}{ss0 \bullet 0}\ \mathrm{A}_{-,2}}{ss0 \bullet 0}\ \mathrm{A}_{0,2}}{s0 \bullet 0}\ \mathrm{A}_{-,1}}{s0 \bullet s0}\ \mathrm{A}_{+,2}}{0 \bullet s0}\ \mathrm{A}_{-,1}}{0 \bullet ss0}\ \mathrm{A}_{+,2}}{0 \bullet ss0}\ \mathrm{A}_{0,1}$$

`Theorem` **2.2.5** *A register machine $M$ with initial state $S$ halts iff the formula $A(T(S)) \rightarrow \langle T_M \rangle \top$ is valid (Where $\top$ is any logical truth).*

Proof: by induction over elementary steps/state transitions. □

We have shown that $G_{\mathcal{L}}$ is undecidable at least for some choices of $\mathcal{L}$. It is also easy to find an $\mathcal{L}$ where $G_{\mathcal{L}}$ will be decidable (take an empty inference system, for example). Decidability for particular interesting logics, such as $G_S$, remains unresolved.

## 2.3 Comparison to other systems

Sentence-based logics of belief are not new; a number of varieties has been proposed over the years. In this section, I briefly discuss these logics and point out where my contribution differs from previous work.

### 2.3.1 Konolige's Deduction model of belief

Konolige (1986) builds his logic primarily as a tool for modeling belief ascription. In this system, each agent $i$ has an associated belief operator $\mathbf{B}_i$.

The state of an agent is modeled as a set of sentences $B_i$ plus a deduction system $\rho(i)$; $\mathbf{B}_i\eta$ is considered true iff $\eta$ is derivable from $B_i$ using $\rho(i)$.

One feature Konolige demands of his agents' deduction systems is that they be *deductively closed*. A system $\rho$ is deductively closed if, given a set of sentences $B$ and sentences $\eta$ and $\zeta$, if $B \vdash_\rho \eta$ and $B, \eta \vdash_\rho \zeta$, then $B \vdash_\rho \zeta$. Such a restriction makes Konolige's system static: an agent's process of reasoning does not change the truth values of any formulas in his logic.

Konolige proves that the logic is both sound and complete with respect to the class of models he considers.

### 2.3.2 Step/Active logics

Step logics (Elgot-Drapkin, 1988) take seriously the idea that an agent's inferential process takes place in time. In step logics, formulas are prefixed with integers, indicating the number of an inferential step, or a moment in (discrete) time. A rule in step logics looks like this:

$$\frac{(i)\phi_1, \ldots \qquad (i)\phi_k}{(i+1)\psi}$$

This means that an agent who believes $\phi_1, \ldots \phi_k$ at time moment $i$ will believe $\psi$ at the next moment.

Such architecture of the logic has the effect that at each moment in time *every* formula that can in principle be derived in one step is actually derived by the next moment. When a certain derivation, represented in tree form, has several branches, the moment at which the conclusion is reached corresponds not to the size of the tree, but its depth.

Moreover, since the logic models an undirected search, in many realistic situations the size of the agent's belief set will grow very rapidly.

Jago (2006, Chapter 4) presents a variant of this framework that can model derivations employing additional assumptions.

### 2.3.3 Duc's dynamic epistemic logic

In his Ph. D. thesis, Duc (2001) proposes to apply the apparatus of dynamic logic to inferential processes. He briefly discusses the suggestion to have a separate action corresponding to each rule used by the agent — essentially, my $F_{\mathcal{L}}$ (pp. 28–30), but quickly dismisses this system as too complicated. He then settles for a single modality $\langle F \rangle$, with $\langle F_i \rangle \alpha$ meaning "$\alpha$ is true after some course of thought of $i$". (In our logic, this modality would be expressed as $\langle \text{Infer} \rangle$.) Duc proves that this simplified system is consistent and embeds normal modal logics into variants of his dynamic epistemic logic.

In Chapter 5 of Duc's dissertation a logic of algorithmic knowledge is discussed, with ${K^n}_i \alpha$ interpreted as "agent $i$ can know $\alpha$ after $n$ steps". In contrast to step logics, there is no assumption that at each next moment in time, all the derivation steps that can be performed are actually performed.

### 2.3.4 Agotnes's system

Ågotnes (2004) has both static and dynamic versions of his system. The static version has operators $\triangle S$ and $\triangledown S$, where $S$ is a term denoting a finite collection of sentences. $\triangle S$ is interpreted as "The agent believes at least the sentences in $S$", while $\triangledown S$ means "The agent believes at most the sentences in $S$". One can postulate that the agent's belief sets are closed under certain inference rules (this leads to a system similar to Konolige's). The static system is proved to be both sound and complete.

In the second part of his thesis, Agotnes extends the system with inference rules. Moreover, he allows the conclusion of a rule to be in another agent's belief box, thereby modeling communication. The conclusion describes an agent's (or agents') state after the rule application. The new state may be smaller than the initial one, thereby modeling belief revision and making the reasoning nonmonotonic. ATL (Alternating Time Logic) is used as a framework for expressing state changes, employing notions of group strategies ensuring certain formulas. The expressive power of the resulting language is immense. Still, at most it is possible to state that an agent who possesses a given set of rules can reach a certain conclusion — eventually or in a given number of steps. It is not possible to express how the rules should be combined together and in what order.

### 2.3.5 Sentential epistemic logics by Jago

Jago (2006) develops a variety of logics where change of state represents inference steps taken by an agent. First, he extends Active logics with ability to handle proofs involving assumptions, obtaining a logic he calls TRL – Timed Reasoning Logic. The mechanism is to enrich the notion of an agent's state: a state is taken to be a collection of contexts, with each context marked by a set of assumptions taken in that context, and containing a set of formulas derived using those assumptions.

Jago also considers a modal logic like that of Duc, with a step of derivation corresponding to a change from state $s$ to an accessible state $t$. Unlike Duc, the basic modality represents just one step, not a whole possible train of derivation. Jago's main interests lie in modeling AI agents, so he makes an assumption, reasonable in that context, that the set of rules accessible to an agent is finite. (In Jago's terminology, a rule is what I call a single instance of a rule). Under this assumption he is able to provide a complete axiomatization for his logic. A multi-agent version of the logic handles communication between agents through a special kind of rules.

In the final chapter of his thesis, Jago combines the rule-based epistemic logic with the TRL's mechanism for handling assumptions. The resulting system is able to model a reasoner that uses natural deduction. The main difference in expressive power compared to my $F_{\mathcal{L}}$ system is that there's still only one modality symbol; only the size of the proof is represented in the epistemic logic's formulas, not any particular shape the derivation might take.

In (Jago, 2009), the epistemic logic for rule-based reasoning is extended to the case of nonmonotonic reasoning through belief revision. Again completeness is proved for the version with a limited number of rules known to an agent.

### 2.3.6 Small-step dynamic epistemic logics by Velázquez-Quesada

In Chapter 2 of his Ph. D. thesis, Velázquez-Quesada (2011) introduces a logic that combines the possible world setup of Dynamic Epistemic Logic (see, e. g., Ditmarsch et al. (2007)) with syntactic information about explicit beliefs of an agent. In contrast to Duc and Jago, Velázquez-Quesada distinguishes inferential actions up to a rule instance. A rule application is formally expressed as a deterministic model-transforming operation, in the style of DEL. Such a fine-grained approach to rules prevents one from identifying inferential steps using different instances of the same rule; thus, it makes no sense to talk about iterated rule applications. On the other hand, the resulting logic is proved to be complete.

In subsequent chapters, Velázquez-Quesada considers related logics employing notions of awareness and belief, the latter being treated as a by-product of a plausibility partial ordering among possible worlds (thus, an agent that believes $p$ can still consider $\neg p$ possible).

### 2.3.7 Logic of proofs

Logic of Proofs, LP (Artemov, 1994) was developed to handle certain problems in proof theory; initially, it was given semantics based on interpretations of arithmetic. Later, however, a different kind of semantics was provided by Mkrtychev (1997) and extended by Fitting (2005), which is closer to the kind of models I consider in this work. Recent publications, such as (Artemov, 2004; Artemov and Nogina, 2005), use the apparatus of Logic of Proofs for studying epistemic logic.

LP deals with proofs in a Hilbert-style system, with a number of axioms and *modus ponens* as the only derivation rule. The role of modality prefixes is played in LP by so called proof polynomials. The polynomials are composed of proof constants and proof variables (constants represent axioms) by means of three operators: unary !, 'proof checker', and binary +, disjunction, and $\cdot$, proof application: $s \cdot t$ produces a proof where the formula $F \to G$, justified by $s$, is applied via MP to $F$, justified by $t$.

As we can see, $\cdot$ closely corresponds to our $\to\mathbf{E}$ in $G_{\mathcal{L}}$ (where the two premises are distinguished from each other), + corresponds to our $\cup$. We have no analogue to the proof checking operator !. On the other hand, LP has no means to express iteration.

## 2.4 Possible extensions

In this section I would like to sketch some ways of extending my logical systems. Detailed investigation of these extended systems is left for future work.

#### 2.4.0.1 Rules for predicate logic

The format of rules described in this chapter, both for $F_{\mathcal{L}}$ and $G_{\mathcal{L}}$, is well suited for systems such as sentential logic and modal logic. Natural deduction systems for predicate logic do not fit this format because certain rules (existential exploitation and universal introduction) contain the additional requirement that a new constant be selected.

This is easily remedied if we consider these a kind of subproof rules (in the Fitch-style system) or rules with an additional premise (in the Gentzen-style system). The subproof 'assumption'/additional 'premise' in this case will not be a formula but the name of the new constant, but similarly to normal premises they have to be discharged before the subproof is finished.

#### 2.4.0.2 Introspection

It is interesting to consider the case where the agent's language itself includes the **B** operator and the dynamic logic modalities. Since our semantics does not make any claim as to whether the agent's beliefs are true or consistent, the inclusion of such operators does not present any particular problem. It makes sense, however, to provide the agent with an introspection ability.

Two kinds of introspection are possible. First, the agent can form beliefs about his own beliefs. A positive introspection rule would look like this:

$$\frac{\eta}{\mathbf{B}\eta}\ \textsc{PosInt}$$

It is not possible to specify a negative introspection rule within our format.

Second, an agent can introspect his own reasoning process. In the Fitch-style system, each simple rule of the form

$$\frac{\zeta_1 \qquad \zeta_2 \qquad \ldots \qquad \zeta_n}{\eta}\ \mathrm{R}$$

can have a corresponding introspection rule

$$\frac{\mathbf{B}\zeta_1 \qquad \mathbf{B}\zeta_2 \qquad \ldots \qquad \mathbf{B}\zeta_n}{\langle \mathrm{R} \rangle \mathbf{B}\eta}\ \mathrm{R}\text{-}\textsc{Int}$$

A subproof rule

$$\begin{array}{c} \zeta \\ \ldots \\ \dfrac{\eta}{\theta}\ \mathrm{Q} \end{array}$$

corresponds to an introspection rule

$$\frac{[\mathbf{O}](\mathbf{B}\zeta \rightarrow \langle \mathrm{Infer} \rangle B\eta)}{\langle \mathrm{Q} \rangle \mathbf{B}\theta}\ \mathrm{Q}\text{-}\textsc{Int}$$

In the Gentzen-style system, for a rule

$$\frac{\Gamma, \zeta_1 \vdash \eta_1 \qquad \ldots \qquad \Gamma, \zeta_k \vdash \eta_k \qquad \Gamma \vdash \eta_{k+1} \qquad \ldots \qquad \Gamma \vdash \eta_m}{\Gamma \vdash \theta}\ \mathrm{R}$$

one can have a corresponding introspection rule

$$\frac{\Gamma \vdash [\mathbf{O}](\mathbf{B}\zeta_1 \rightarrow \langle \alpha_1 \rangle B\eta_1) \qquad \ldots \qquad \Gamma \vdash \langle \alpha_{k+1} \rangle \mathbf{B}\eta_{k+1} \qquad \ldots}{\Gamma \vdash \langle \frac{\alpha_1 \ldots \alpha_{k+1} \ldots}{} \mathrm{R} \rangle \mathbf{B}\theta}\ \mathrm{R}\text{-}\textsc{Int}$$

Combined with rules for dynamic logic, each possible proof in $F_{\mathcal{L}}$ or $G_{\mathcal{L}}$ may be reflected within the agent's reasoning process.

Extra care should be taken to avoid paradoxical states — such as an agent simultaneously believing $\eta$ and $\neg\mathbf{B}\eta$. Introspection should thus probably be combined with non-monotonicity and some kind of 'repair' rules, which fire obligatorily upon acquiring any belief (see the next subsection).

#### 2.4.0.3 Algorithmic derivation

The system as it is set up so far only represents what is *possible* for the agent to derive. Every action in our model is reflexive; the agent can always perform derivations on those parts of his knowledge that do not concern us. He has, so to say, no obligation to work for us. No statement of the form $[\alpha]B\eta$ is going to be true in a state that does not already have $\eta$.

One way to remedy this is to introduce actions that work deterministically. Some may consist in application of all possible instances of certain rules — such as simplifying all the conjunctions in the current belief state. Others might include more complicated goal-directed behaviour; but a sensible specification of the goals is likely to make the logic much more complex than it already is.

A more flexible approach is to enrich the language with tests, as in standard PDL. Such tests may check whether a certain formula is believed or whether a rule is non-vacuously applicable. Tests will allow us to build programs as actions. For example, the action of performing all possible conjunction simplifications would be

$$\wedge\mathbf{E}^*; \sim\wedge\mathbf{E}?$$

where the test action $\sim \wedge\mathbf{E}$? only succeeds when the rule $\wedge\mathbf{E}$ is not applicable.

#### 2.4.0.4 Belief revision. Multi-agent logic. Communication

These topics are considered together because in all three cases I can simply adopt solutions proposed in (Ågotnes, 2004; Jago, 2006).

First, we can drop the monotonicity requirement for the rules we use and introduce actions that drop certain beliefs. A simple form of such actions, discussed in (Jago, 2009), is

$$\frac{\eta_1 \qquad \ldots \qquad \eta_k}{\sim \zeta}\ \mathrm{R}$$

That is, given beliefs $\eta_1, \ldots \eta_k$ are present in the agent's belief box, the agent is allowed to drop his belief in $\zeta$.

Ågotnes (2004) considers a more general notion of a mechanism for belief change; rules are just one way of specifying such a mechanism. A rule consists of a specification of the *complete* belief state of an agent before the rule is applied, together with a specification of the state after its application. Thus, a monotonic *modus ponens* rule looks like this:

$$\frac{t \sqcup \{p, p \rightarrow q\}}{t \sqcup \{p, p \rightarrow q, q\}}$$

while a rule that allows the agent to forget the premises is specified by

$$\frac{t \sqcup \{p, p \rightarrow q\}}{t \sqcup \{q\}}$$

(In Agotnes's semantics, an agent's state after execution of the rule has to be *at or above* the state described by its conclusion.)

Switching from a single-agent setting to a multi-agent one is a matter of multiplying the belief operators **B** and every rule action by the number of agents. A model for the $n$-agent system will be the $n$-th power of the original model; each state is an $n$-tuple of finite sets of sentences.

Providing agents with the ability to communicate is slightly more challenging. Following Jago (2006), we can add to the agents' internal language two modalities, $\textcircled{?}_{ij}$ ‘ask' and $\textcircled{\vdash}_{ij}$ ‘tell'. Whenever a formula headed by such a modality appears as the conclusion of an agent $i$'s rule, the execution of such a rule puts the formula in both agent $i$'s and agent $j$'s belief boxes. Informally, $\textcircled{?}_{ij}\eta$ means ‘Agent $i$ asks agent $j$ whether $\eta$ is true, and $\textcircled{\vdash}_{ij}\eta$ means ‘Agent $i$ tells agent $j$ that $\eta$'. For such interpretations to be sensible, the agents must have corresponding rules. For example, this is a pair of rules for answering questions truthfully:

$$\frac{\textcircled{?}_{ji}\eta \qquad \eta}{\textcircled{\vdash}_{ij}\eta}\ \mathrm{QA^{+}}_{ij} \qquad \frac{\textcircled{?}_{ji}\eta \qquad \neg\eta}{\textcircled{\vdash}_{ij}\neg\eta}\ \mathrm{QA^{-}}_{ij}$$

A more general approach, taken from Ågotnes (2004), is to have double indices on rules, so that the premises of the rule have to be satisfied in the belief state of one agent, and the conclusion is added to the belief state of (possibly) another agent. For example, here is a rule that allows agent $i$ to inform agent $j$ of an arbitrary belief he has:

$$\frac{t \sqcup \{p\}}{\{p\}}\ \mathrm{C}_{ij}$$

In either case, a question arises whether actions of agents take place in an arbitrary order, in some predefined order, or simultaneously. The first alternative does not require any change to our semantics for the logic; the second can be expressed as a restriction on the form of admissible actions. The third one requires combined elementary actions: each elementary action becomes a tuple of $n$ actions undertaken by the $n$ agents present in the model.

## 2.5 Conclusion

In this chapter, we introduced a family of logical languages and investigated their properties. We found that these logics allow us to specify belief states of agents equipped with syntactic reasoning capabilities that are inexpressible in preceding systems with similar semantics. In the following chapters we will put this additional expressive power to work.

The more powerful of our logics, $G_{\mathcal{L}}$, is in the general case undecidable.

While logic $G_{\mathcal{L}}$ seems more adequate for capturing shapes of inference, it is also considerably more cumbersome from the notation point of view than $F_{\mathcal{L}}$. That is why in the subsequent chapters, where I discuss applications of my theory, I stick to the simpler language of Section 2.1.

# Chapter 3

# Indirect speech

Indirect speech is perhaps the clearest example of a construction where one cannot take the meaning of its sentential complement to be just a set of possible worlds. At the same time, unlike direct quotation, it is not possible to maintain the position that the complement is just a string of symbols. The string of symbols — the way the original utterance being cited was phrased, — plays a role in the truth conditions of an indirect speech report, but indirect speech does not require the original utterance to be reproduced in its entirety. It is my hope that a logic of the kind introduced in Chapter 2 can serve as a tool that allows us to link the content of the report to the content of the utterance being reported.

The question to be resolved is: what is the relation between the original sentence $p$ uttered by some agent $A$ and the report $A$ *says that* $q$? What $q$'s can be based on a given $p$ and what $p$'s can serve as the source of a given $q$? (I remind the reader that change of indexicals' point of reference is not my primary area of interest, so I ignore it.)

Of all the linguistic constructions to be discussed in this thesis, indirect speech admits the narrowest range of operations that stand between $p$ and $q$. For this reason, I am going to use this chapter as a model for the later ones. I will simplify the treatment as much as possible in order to expose the structure of the argument.

Section 3.1 tries to restrict the range of data I am working with so that the task becomes more manageable. Section 3.2 surveys some of the previous work that is concerned with the same problems I am dealing with. Section 3.3 lists the inferential operations that can be applied to a sentence uttered by the original speaker in producing an instance of indirect speech. In Section 3.4 a preliminary formal analysis is built based on observations of Section 3.3. A more adequate analysis of indirect speech will be presented in §4.4.

## 3.1 Limiting the data

Before we can start working on the solution, we need to delimit the range of data we are dealing with. There are two classes of usage that we'd like to exclude from main consideration:

- sloppy usage;
- *de re* indirect speech.

Furthermore, in this chapter I only consider sentences expressible in first order predicate logic as the content of reported utterances. Later, in §4.4, this limitation will be lifted.

### 3.1.1 Sloppy usage

Differences in acceptability judgments among speakers and among contexts of utterance exist for almost any linguistic construction, but in the case of indirect speech (as well as for belief reports and hearsay evidentials), the gray area where the ascription in question does not seem definitely false, but rather slightly awkward, is particularly wide. While there are pairs of original utterances and reports that everyone will agree with (John, the primary speaker, says: *It's raining.* I report: *John says that it's raining*), and pairs everyone would consider inadmissible (John says: *It's raining.* I report: *John says that Quito is the capital of Ecuador*), there are also many intermediate cases. One example: at a party, John says: *I am going home*; I report: *John says that he is leaving.*

I dislike being in a position of a language lawyer, but in order to work with such a blurry distinction one has to sharpen it a bit. I intend to be as strict as possible in my judgments, only allowing those sentences that are absolutely unproblematic. In particular, I will employ the following test:

Test **3.1.1** *The pair $(p, q)$ is unacceptable if an informed observer can issue a correction "A did not say that q, he said that $q'$" (for some $q'$ that is closer to p than q).*

### 3.1.2 Indirect speech *de re*

Indirect speech *de re* can be treated as anaphoric (3) or bound-variable (4) links connecting the prejacent of indirect speech with material outside it.

(3) Peter$_i$ went away. John said that *he*$_i$ wasn't feeling well.

(4) Every girl$_i$ suspected that John said he likes *her*$_i$ .

In some works (Soames, 1989; Brasoveanu and Farkas, 2007) it is claimed that replacing names and descriptions with coreferential names and descriptions is allowed in indirect speech. Unfortunately, in the case of indirect speech it is not possible to distinguish a *de re* report from a *de dicto* report with such a replacement, so I don't include this operation in the list of what's allowed in indirect speech. For belief reports, I do think it is possible to derive *de dicto* belief ascriptions with coreferential NP replacement; see p. 78. Such a confusion, however, makes it worthwhile to discuss *de re* indirect speech reports in some detail.

In the cases of anaphora, pronouns can be used not just for nominal constituents, but for verbs as well.

(5) John says that Bill [works too much]$_i$. Bill says *the same*$_i$ about John.

As for the bound variable cases, the binding operator (definite or quantifier) may stay in place in the surface syntax (i. e., if we adopt the quantifier raising theory, the raising is covert).

(6) John: Michael will come to the party.
Peter: John says that *a friend of his* will come to the party.

Using a theory of presuppositions in the vein of van der Sandt (1992), we can even handle *de re* uses of proper names. A proper name has a binding operator as a presupposition that rises to the outermost level of the sentence (or even discourse), crossing the indirect speech border. A variable is left in place, to be bound when the operator in the presuppositional part of the meaning is accommodated.

(7) John: My favourite writer is Mark Twain.
Peter: John said that his favourite writer is Samuel Clemens.

A restriction has to be placed on this kind of name change that arises from the strict approach to truth conditions of indirect reports outlined in the previous section: it is presupposed that the identity of referents referred to by the names and/or definite descriptions used in the primary and the secondary utterances is common knowledge among all the participants (speaker and audience) of both speech acts. Otherwise the indirect speech report is open to the criticism: *John did not say that his favourite writer was Clemens, he said it was Mark Twain!* One can either treat examples like that as true but misleading (which gives rise to a wide range of theories of *de re* speech and propositional attitude attribution), or just declare them infelicitous, failing the presupposition.

Apart from definite descriptions and proper names, the only type of covert quantifier raising allowed in such cases is specific indefinites. They are specific because the secondary speaker knows the identity of the object (it was referred to in the primary utterance).

It is a well known property of specific indefinites (Fodor and Sag, 1982) that they don't obey island restrictions on scope; for example, indefinites occurring in the antecedent of a conditional can have scope over the whole conditional:

(8) John: If Michael comes to the party, there may be a fight.
Peter: John said that if *a friend of his* comes to the party, there may be a fight.

Other kinds of covert quantifier raising are disallowed:

(9) *It is known to everyone involved that John has exactly three students: Bob, Bill and Mary*
John: Bob, Bill and Mary are sleepy today.
Peter: #John said that every student in his class is sleepy today.

Similarly to cases of failed common knowledge of identity, one could perhaps argue that quantificational *de re* indirect speech ascription in examples like (9) is true but misleading. I can't think of any test that would distinguish between these two analyses; however the theory that such utterances are true would carry an additional burden of explaining the intuitive false judgments, so I prefer simply to treat the sentences as false.

It is not always a matter of choice for the secondary speaker whether to use *de dicto* or *de re* indirect speech reports. *De re* reporting may be inevitable in the following cases:

- Report is quantifying over a number of primary utterances, as in (4).
- A proper name or description used in the primary utterance is not familiar to the secondary addressee.
- Politeness. Sometimes replacing the proper name or pronoun can be dictated by contextual factors such as official or informal situation or hierarchical relations between participants of conversation.

(10) John: I talked to Josh today.
Peter: John says he talked to Prof. Dever today.

(11) Russian: Ivan, to Vasiliy: *Petr **tebja** včera videl*
'Peter saw **you** (informal) yesterday.'
Ivan, to Vasiliy, in a more formal setting: *Ja skazal, čto Petr včera **Vas** videl*'
'I said that Peter saw **you** (polite) yesterday.'

In such examples, shift occurs between linguistic registers within the same language, and, as such, it has a lot in common with the shift between different languages (§3.3.1). On the other hand, the type of the situation (formal or informal) can be considered a special kind of indexical, although, obviously, it is absent from the standard Kaplanian list of speaker, place, and time.

- The reporter may not have full information about the original utterance; for example, when he has not heard or understood all of it.

  (12) A: John … *(mumble)*… the vase and everyone's upset.
  B: A. says that John did something to the vase.

One problem that arises in analyzing indirect speech *de re* is the existential commitment of wide scope quantifiers[1]. Let's assume Michael is paranoid, and thinks there is a maniac called Smith who is trying to kill him. In such a situation, report (13) is felicitous:

(13) Michael: Smith is hiding under the table!
John: Michael says that a maniac is hiding under the table.

Under the wide scope quantification analysis, John would have to subscribe to the existence of Smith in uttering his report, which he intuitively does not.

One could attempt to solve this by allowing replacement of coreferential NPs in indirect speech after all, the way I do for beliefs, and treat the example as a *de dicto* report. But a slight twist in the example shows that this move fails: let us now assume that Smith is not a figment of Michael's imagination: he is a humble and harmless co-worker who Michael wrongly suspects of being after him. In such a setup, (13) seems much less appropriate, even though

[1] Pointed out to me by Josh Dever (p. c.).

in Michael's belief worlds, which are to be taken into account with a narrow scope quantification, Smith *is* a maniac. Note also that, unlike indirect speech report, belief report would be felicitous in this case.

Indirect speech *de re* is not the main focus of our investigation, so, having noted its existence, we concentrate on indirect speech *de dicto*.

## 3.2 Some previous work

Work investigating relations between the text of a primary utterance and the corresponding indirect speech includes Soames (1989); Cappelen and Lepore (1997); Brasoveanu and Farkas (2007).

Soames (1989) indicates that many propositional attitude verbs support conjunction simplification [p. 396]. (He includes 'say' in this list.)

He also [p. 402] claims that propositional attitude reports preserve truth under change of proper names. Soames uses this to motivate his theory of structured (Russelian) propositions. He does see the problems that arise from free use of substitution. In order to dismiss those problems, he needs to state that "semantic facts of English are not always fully accessible to simple introspection by competent speakers" [p. 419]. Since I am trying hard to avoid just such a claim, I have to specify that certain background conditions are to be met in order for substitution to work (see p. 47). Therefore Russelian propositions are not an acceptable alternative for me. I will admit change of coreferential proper names and descriptions as an acceptable operation for belief reports in Chapter 4.

In his analysis of assertion [p. 411] Soames uses the phrase *can be readily inferred*:

> "$x$ asserts that $S$" characterizes the agent as having assertively uttered a sentence $S'$ in an associated context $C'$, such that for some $S''$ that can be readily inferred from $S'$, the content of $S''$ in $C'$ = the content of $S$ in the context of the report.

He never specifies what counts as such *ready inference*, though.

Cappelen and Lepore (1997) argue that the actual practice of speech reporting is too unsystematic and context dependent to give any precise rules. My approach will in fact disallow many of the examples they provide. Reports that I do recognize as acceptable are those that will work in every context, and survive the challenge test, as well as their test with attaching "literally" to the report. Their work mentions most of the classes of inference that I am going to discuss: conjunction simplification, dropping of adjectives and adverbs, coreferential proper name substitution.

Brasoveanu and Farkas (2007) contrast indirect speech reports with belief ascriptions. They emphasize the fact that, unlike belief reports, indirect speech is always anaphoric to some particular utterance. Another condition they claim applies to indirect speech is *faithfulness to the meaning dimensions*: the at-issue entailments of the complement clause in an indirect speech report should be a subset of the at-issue entailments of the primary utterance; same for implicatures, and the presupposition/at-issue content division of the source speech act must be preserved in the report. Among the examples provided by Brasoveanu and Farkas we see conjunction simplification, dropping of some verb arguments (including passivization that drops the original subject), replacing coreferential names and descriptions, certain lexical inferences, introduction of presupposition-carrying operators, provided the presupposition is known to be satisfied.

(14) Sam: Mary visited Santa Cruz last week.
John: Jane was in Santa Cruz last week and Sam says Mary was here too.

(here, *visit L* implies *be at L*; the presupposition-carrying *too* is introduced.)

(15) Sam: Sue wants a Porsche.
John: Sam says that Sue wants a car.

Such examples would be disallowed under the strict conditions of Test 3.1.1.

## 3.3 Inferences allowed in indirect speech

My goal in this section is to build an exhaustive list of operations that can be applied to the text of a primary utterance when forming an instance of indirect speech while meeting the demands of Test 3.1.1. The list clearly depends on the somewhat arbitrary details of the task set-up; slight variations, such as incorporating indexicals into the picture, can result in the expansion of the list.

In forming $q$ (in the *de dicto* indirect speech expression *A said that q*) on the basis of an utterance $p$, the following operations can be applied to $p$:

- Translation between languages;
- Conjunction simplification;

These operations guarantee that $p$ entails $q$.

### 3.3.1 Translation between languages

The prejacent of indirect speech is always expressed in the same language as the rest of the sentence (and normally, the rest of the text). If the primary utterance to which a particular use of indirect speech refers is in fact made in some other language, it needs to be translated.

(16) Ivan: Dvaždy dva četyre. *(Two times two is four — Russian)*
Peter: Ivan says that two times two is four.

Sometimes this requirement is violated, and in such cases we can speak about a mixture of indirect speech and direct quotation (Maier, 2007; Shan, 2010).

(17) Bush said that the terrorists *misunderestimated* him.

Such a mixture of the primary speaker's idiolect with the one of the secondary speaker is possible only when they are close enough to be mutually understandable.

### 3.3.2 Conjunction simplification

Several operations can be interpreted as subspecies of conjunction simplification[2]:

- Simple conjunction;
- Dropping veridical adverbs and intersective adjectives;
- Dropping arguments.

[2] For the moment I am ignoring the fact that the conjunctions in question may occur in the scope of other operators. See §4.3.

First, there are conjunctions as such: sentential (18) or constituent (19, 20):

(18) John: Quito is the capital of Ecuador and La Paz is the capital of Bolivia.
Peter: John says that La Paz is the capital of Bolivia.

(19) John: Michael took off his trousers and went to bed.
Peter: John said that Michael went to bed.

(20) John: Michael drinks both beer and wine.
Peter: John says that Michael drinks beer.

Note that it is not the presence of syntactic conjunction as such that allows simplification, but its interpretation as logical conjunction (rather than an operation creating sum individuals)

(21) John: Michael, Bill and Bob carried the piano upstairs.
Peter: $^{\#}$John says that Michael carried the piano upstairs.

Second, intersective adjectives, as well as adverbs of time, place and manner, can be dropped.

(22) John: Michael bought a red car.
Peter: John says that Michael bought a car.

(23) John: I read the paper thoroughly yesterday at home.
Peter: John says that he read the paper.

All these types of adverbs are veridical — that is, a sentence containing the adverbs implies the sentence with adverbs dropped.

Third, if a verb has an argument that can be dropped without change in interpretation of the other arguments, this drop is allowed in indirect speech.

(24) John: I ate a slice of pizza.
Peter: John says that he ate.

Dropping an argument is allowed only when there are two subcategorization frames for the verb — one containing the argument in question, the other lacking; if the argument is normally obligatory, but can be omitted in some specific circumstances (*He hit, and hit, and hit*), in indirect speech this omission is prohibited.[3]

(25) John: Michael hit Bill.
Peter: $^{\#}$John says that Michael hit.

On the other hand, if the theta-role of (at least one) of the remaining arguments changes between subcategorization frames, changing the frame is, again, prohibited.

(26) John: Michael broke the window.
Peter: $^{\#}$John says that Michael broke.
Peter: $^{\#}$John says that the window broke.

Let us adopt a neo-Davidsonian representation where every verb has an event argument, and both syntactic arguments and adverbs are represented by means of predicates corresponding to theta-roles of the arguments and to adverbs. Thus, the sentence

[3] Nick Asher, p. c.

(27) Michael ate a sausage slowly.

would be represented as

$$\exists e.\exists x.(\mathbf{sausage}(x) \wedge \mathbf{eat}(e) \wedge \mathrm{Agent}(e) = \mathbf{m} \wedge \mathrm{Patient}(e) = x \wedge \mathbf{slow}(e))$$

It is easy to see that under such a representation all the operations considered in this subsection become special cases of conjunction simplification.

Nick Asher (p. c.) suggests that the fact only non-obligatorily expressed arguments are allowed to be dropped is evidence that representation in the style of original Davidson's account is preferable which equips the verb with an event variable, but keeps all the arguments together:

(28) Michael hit Bill.
$\exists e.\mathbf{hit}(e, \mathbf{m}, \mathbf{b})$

In my opinion, the evidence is inconclusive, since the impossibility of expressions like *Michael hit* can be caused by purely syntactic restrictions.

Note that three of the mentioned prohibited cases: conjunction simplification when conjunction denotes a sum individual, nonveridical adverbs, and argument dropping when it leads to change in the theta-roles of the remaining arguments, do not preserve entailment.

## 3.4 Formal presentation

We now try to characterize the inference rules one can use in forming the prejacent of an indirect speech construction. Unfortunately, there is no way to do this for translation into another language without complicating the system considerably. On the other hand, if we assume neo-Davidsonian representation of the prejacent, three kinds of simplification (conjunction simplification, dropping veridical adverbs and dropping arguments) can be treated as one rule:

$$\frac{\phi \wedge \psi}{\phi} \wedge\mathbf{E}$$

Rules used in indirect reports will typically have to be applied within the scope of one or more operators such as quantifiers or conditionals; moreover, applicability of the rules depends on the context where the premise is located within the formula. I delay discussion until §4.4; in that section, I will also be able to handle a much wider fragment of the language (including, for example, subsective adjectives and focus operators).

To capture the meaning of indirect speech, we need, first, to introduce quantification over utterances (over sentences with resolved indexicals) — since indirect speech is only true when a primary utterance exists, but does not allow us to reconstruct that utterance completely. Second, we need a way to specify that a certain sentence/formula is derivable *given* a certain premise. In building a translation of indirect speech into formal language, I use modal necessity — if one explicitly believes $p$, one will always be able to derive $q$ using certain inference rules. With these details taken into account, I give the following semantics to indirect speech:

$$[\![A\ \textit{says that}\ q]\!] = \exists p\ (\mathbf{say}(A,p) \wedge \Box(\mathbf{B}p \rightarrow \langle \wedge\mathbf{E}^{*} \rangle \mathbf{B}q))$$

## 3.5 Concluding remarks

In this chapter, I have applied my general framework to build a model that can predict certain entailment relations between sentences uttered by a primary speaker and sentential complements of indirect speech constructions. This account is deliberately narrow in scope: whenever it predicts an indirect speech expression is allowed, it should be allowed irrespective of contextual factors; but many instances of indirect speech that actually occur are not predicted.

As soon as we lift the artificial restriction that the contents of the reported utterance be expressible in the first order predicate logic, more classes of operations on those contents becomes evident. One is dropping focus operators like *even*, *only*, clefts and pseudo-clefts.

(29) John: It was George who ate the watermelon.
Peter: John says that it was George who ate the watermelon.
*or*
Peter: John says that George ate the watermelon.

(30) John: What George did was eat the watermelon.
Peter: John says that what George did was eat the watermelon.
*or*
Peter: John says that George ate the watermelon.

(31) John: Only Michael solved the problem.
Peter: John says that Michael solved the problem.

(32) John: Even Michael solved the problem.
Peter: John says that Michael solved the problem.

Note that in (31) the complement of indirect speech is a *presupposition* of the original utterance (and even the fact that there is such a presupposition is being debated, see e. g. (Horn, 1996; Beaver and Clark, 2008) for contrasting viewpoints). This presents a problem for the account of Brasoveanu and Farkas

(2007), which requires the presupposition/at issue distinction of the original utterance to be retained in the indirect speech.

Similarly, once the restriction on utterance content is dropped, we no longer need to restrict ourselves to intersective adjectives, but we still need the adjective to be subsective:

(33) John: Michael is a good surgeon.
Peter: John says that Michael is a surgeon.

(34) John: Michael offered shelter to an alleged criminal.
Peter: #John says that Michael offered shelter to a criminal.

Note that an inference with a non-subsective adjective would fail to be truth-preserving.

This chapter concentrated on the verb *to say*; it is a question for future investigation whether other verbs of speech behave in the same way and whether they admit the same inferences. In particular, the verb *tell* is worth studying, as well as verbs of assertion like *claim* and *assert* itself.

Another topic where the approach of this chapter has to be extended is indirect questions (and the verb *to ask*). In order to describe which transformations are allowed there, we need to speak about entailment relations between questions (Groenendijk and Stokhof, 1988) and inferences performed on questions.

This chapter is mostly based on my own acceptability judgments. While intuitions with respect to truth conditions of indirect speech are inevitably vague, a quantitative study may put them on a somewhat firmer ground.

# Chapter 4

# Belief ascriptions

This chapter deals with belief ascriptions. Our task for analyzing this type of constructions is: what sentences $p$ does an agent $A$ need to have in his belief box in order for the ascription $A$ *believes that* $q$ to count as true?

This kind of task is somewhat more problematic than it is in the case of indirect speech since there the primary utterance $p$ is always, at least in principle, publicly accessible. Explicit beliefs — sentences stored in the speaker's mind, — that play the same role in belief ascriptions are not public and typically it is not as easy to determine whether a given belief is present as a stored Mentalese sentence in one's mind — even for the agent himself. The following circumstances will be considered reliable enough evidence that (translation of) $p$ is explicitly present in $A$'s belief box:

- Sincere assertion of $p$ by $A$. We assume that we can determine whether a given assertion is sincere.

- $A$ hearing an assertion of $p$ and accepting it. Again, it is assumed that we can recognize when $A$ accepts $p$ and distinguish it from cases where he disbelieves it but does not bother arguing against it, etc.

- $A$ explicitly believing $q$ such that $p$ is a presupposition of $q$:

(35) John: Bill stopped smoking.
$\therefore$ John believes that Bill used to smoke.

In real life, of course, there are many more cases where a certain explicit belief is ascribed. Suppose Bill is John's best friend, John normally uses the name 'Bill' to refer to Bill, and Bill stands right before John, with nothing obscuring the view. It seems reasonable to ascribe to John the explicit belief *I see Bill.* However even in such cases it is hard to determine exactly which of a number of non-equivalent internal sentences John possesses (maybe it's *Bill is standing before me* or even *Bill looks tired today*), so we will try to avoid such cases in our examples.

Similarly to indirect speech, there is a large amount of sloppiness in ascribing beliefs, and that leads to a wide gray area where it is not clear whether a given ascription is warranted. Here, as there, we might want to distinguish between false ascriptions and ascriptions that are technically true but misleading. Again, as in the case of indirect speech, I take the narrowest possible position, applying the following test:

`Test` **4.0.1** *Explicit belief in p by an agent A serves as a justification for the belief ascription of q* only *if "A explicitly believes that p but does not believe that q" is a contradiction according to our pre-theoretical intuitions.*

Again, as in the case of indirect speech, we need to acknowledge the existence of *de re* belief ascriptions, even if we do this only to indicate that they are not the primary object of our attention. Rules for forming such ascriptions are the same as in the indirect speech case. Moreover, in English propositional attitude reports can have a syntactic frame unavailable in indirect speech and specialized for *de re* (and *de se*, if one takes this kind of belief ascriptions to form a separate class) ascriptions:

(36) John believes (wants, etc.) himself to be smart.

(37) $^*$John says himself to be smart.

This chapter starts by reviewing various approaches to analyzing belief in Section 4.1. Section 4.2 presents a list of the inferential operations that belief ascriptions support. Some of those operations are sensitive to monotonicity features of their enclosing environment, as discussed in Section 4.3, which introduces a variant of a system called Natural Logic. This, in turn, allows me to improve my presentation of the indirect speech construction. I do this in Section 4.4. Section 4.5 examines the remaining problems with my account. Section 4.6 is a list of examples collected from previous works on belief ascription. I use these examples to test my theory. Finally, Section 4.7 makes my account more formal using the logical apparatus of the kind discussed in Chapter 2.

## 4.1 Theories of belief and belief ascription

Up to this point we spoke as if it were a done deal that human minds operate on sentence-like representations in some internal language and that the right way to approach belief ascriptions is by investigating the properties of those representations. In fact many philosophers disagree with this picture. In this section we survey the alternative viewpoints.

### 4.1.1 Disposition to assent

One approach takes belief to be a disposition to certain behaviour; the easiest such behaviour that indicates belief is verbal. This leads us to the kind of analysis presented in (Carnap, 1947):

> It seems that the sentence 'John believes that D' in $S$ can be interpreted by the following semantical sentence:
>
> > 'There is a sentence $\mathfrak{S}_i$ in a semantical system $S'$ such that (a) $\mathfrak{S}_i$ in $S'$ is intensionally isomorphic to 'D' in $S$ and (b) John is disposed to an affirmative response to $\mathfrak{S}_i$ as a sentence of $S'$.'
>
> (p. 61–62)

Carnap's analysis is complicated by the fact that John might speak a different language from the language of the reporter, but essentially belief in 'D' is ascribed to John iff he is disposed to respond affirmatively to 'D'.

One who takes this position will be led to some particular answers to related questions. First, the object of belief, if there is any, turns out to be a natural language sentence; that is, beliefs are subject to extremely fine individuation conditions. Second, this analysis in itself predicts no systematic correlation (a) between beliefs in different sentences, and (b) between belief in a certain sentence and extra-linguistic behaviour of the believer. Explaining all such systematicity would presumably be a task for biology, neurophysiology, psychology or some other empirical science, not philosophy.

Beliefs are central to epistemology and philosophy of language. Presumably these areas should enjoy some independence from the latest findings in psychology or neurophysiology. Moreover, science in general is concerned with obtaining knowledge, and knowledge is typically considered a kind of belief. A logical circle arises (albeit a rather large one, so it may be acceptable to some).

There is a smaller circularity. Dispositions to a certain response only count given a background assumption of the speaker's sincerity. But what is it to be sincere? One answer immediately comes to mind: $A$ is sincere iff $A$ believes what he says. Unless an alternative account is produced, the sincerity condition itself depends on the analysis of belief.

Even if one does not take disposition to assent to be a proper analysis of $A$ *believes that* $p$, one can regard such a disposition as a reliable *test* whether $A$ believes $p$. That is, instead of constitutive role, we ascribe evidential role to dispositions to verbal behaviour. For example, in (Kripke, 1979) we find the 'strengthened disquotational principle':

> A normal English speaker who is not reticent will be disposed to sincere reflective assent to '$p$' if and only if he believes that '$p$'.

Counterexamples have been proposed to this principle. They concern the 'only if' direction of the biconditional — incidentally, the direction that Kripke himself took to constitute the weaker disquotational principle[1]. One is due to Powers (1978) and concerns puzzles — cases where the answer is 'obvious' once shown to the subject answering the puzzle, but hard to find otherwise. (Powers's own example is 'Is there a word consisting of 4 letters

[1] 'Misdisquotations' from (Moore, 1999) can serve as counterexamples in both directions. But these examples depend on cases of confused identity. In this work I make the simplifying assumption that no such confusion arises.

that ends in E-N-Y?' with the answer 'DENY'). Let $p$ be the puzzle and $q$ the answer to it. Then, in the normal case, an agent has a disposition to answer 'yes' to the question 'Is it true that $q \wedge p$ holds?' ('Is it true that DENY is a word of 4 letters ending in E-N-Y, and there is a word of 4 letters that ends in E-N-Y?'), but not to 'Is it true that $p$ holds?'. Thus, according to the disquotation principle, we would need to ascribe to the subject belief in $q \wedge p$, but not in $p$. This runs counter to many people's intuitions (although Powers himself accepts this as a result).

Another counterexample was presented by Audi (1982). An excited speaker is telling a story to his table companions. If he were asked whether his voice is too loud, he would immediately realize that it is. Thus, using the disquotational principle, we would have to assign to him the belief *My voice is too loud*; this result seems counterintuitive.

Here's one more problematic case. An opinionated man (Lycan, 1985) never admits that he does not know the answer to a question; instead, he says "yes" or "no" based on some information unrelated to the subject matter of the question (for example, on the output of a random number generator). A quick sincere assent of the opinionated man cannot serve as an indication of prior belief.

As for the 'if' direction (if a normal English speaker believes 'p' and is not reticent, he will be disposed to affirm 'p' on reflection), I consider it a good test, a necessary condition for ascribing belief. I will use it to argue against certain theories of belief. (Of course, one might instead accept these theories and reject the principle.) Violations of this principle have a somewhat Freudian flavour: we know you are disposed to deny the truth of $p$ ('I should marry my mother'), and yet this is what we claim you believe (and perhaps the very intensity of your denial serves as additional evidence that you do in fact believe $p$).

### 4.1.2 Interpretationism

One would like to use the whole gamut of behavioural dispositions of an agent, not limited to linguistic ones, in ascribing beliefs to that agent. However such an approach demands a holistic viewpoint, since no belief by itself determines what the agent is likely to do; it does so only relative to other beliefs and desires.

This leads to the following analysis of belief and desire:

> To desire that $P$ is to be disposed to act in ways that would tend to bring about that $P$ in a world where one's beliefs, whatever they are, were true. To believe that $P$, is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which $P$ (together with one's other beliefs) were true. (Stalnaker (1984), p. 15)

Stalnaker (1984) argues that functionalism (defined by him as the view according to which a mental state is individuated by its actual and potential causal relations to stimuli, behaviour and other mental states) inevitably leads to an extremely coarse grained individuation conditions for belief. In his opinion, given an agent's desires, his actions and dispositions to act can only distinguish beliefs up to their truth conditions. Therefore, a belief state can be represented by a set of worlds that the agent regards as (epistemically) possible. A belief ascription would then indicate that in every world of the $A$'s belief state $p$ holds. Thus the object of a belief verb is a set of possible worlds.

Taking belief states to be model-theoretic constructions such as sets of possible worlds has the advantage of simplicity: the meaning of other constructions (such as modals, intensional adjectives and verbs[2]) is analyzed using the same basic machinery, so why not apply it to propositional attitudes as well? Moreover, modal logic is a highly developed and well understood discipline. It would be beneficial to be able to apply its results to the analysis of belief states.

Fischer (1985) considers Stalnaker's arguments for coarse-grained analysis of belief states and concludes that it is not in fact entailed by the functionalist approach. He shows that Stalnaker's argument either uses an implausible formulation of the functionalist definition of belief, or, when a more intuitively correct definition is used, involves quantifying into opaque contexts as well as circularity.

Another consideration makes the combination of functionalism and coarse grained individuation of beliefs implausible. If $A$ is able to speak and desires to be truthful, it is possible that he would assent to a sentence $p$, but would fail to assent to another sentence $q$ with the same truth conditions. Thus, it would seem, $A$ is disposed to act in such a way as to fulfil his desires provided that $p$ is true, but not provided $q$ is true. Either we should individuate beliefs finely enough to distinguish the proposition expressed by $p$ and one expressed by $q$, or ascribe irrationality to $A$. As soon as we include verbal behaviour as one of the components that determine belief individuation, it is hard to maintain the coarse grained position.

Moreover, the most straightforward application of an approach based

[2] At least some of these, such as *search*, should be analyzed in the same way as propositional attitudes.

on sets of possible worlds leads to counterintuitive results, namely the logical omniscience problem. An agent that believes a proposition $p$ thereby believes all its logical consequences. In particular, necessary truths, such as true mathematical statements, are believed by all agents. It is also impossible to have different attitudes to two logically equivalent statements. Clearly agents in the real world do not have these characteristics.[3]

Several strategies have been attempted in order to overcome this difficulty. First, it can be claimed that the possible world analysis gives us the notion of 'implicit belief'. But implicit belief seems to play no role in explaining and predicting the agents' behaviour (at most, it can play a normative role, telling us how agents *should* act). In order to be useful for these purposes, the implicit belief theory should be supplemented, if not replaced, by an account of explicit belief, and what the natural language expression *to believe* denotes is much closer to explicit than implicit belief.

To avoid closure under logical consequence, one can claim that an agent's belief state consists of multiple substates. Each substate is a set of possible worlds (and thus beliefs in this substate are closed under logical consequence), but the substates may be mutually incompatible, and the process of reasoning is required in order to combine them. (This is the position advocated by Stalnaker (1984). A logic based on this 'society of minds' idea is presented in Fagin and Halpern (1988)). However, in order to prevent logical consequences from being believed, one would have to posit an indefinite number of such substates. In particular, should we place every axiom of a mathematical theory in a separate substate? And, assuming some theory is

[3] Whitsey (2003) is an excellent survey of various approaches to logical omniscience. See also Fagin et al. (2003, Chapter 9)

finitely axiomatizable, once an agent deduces a theorem that uses all of the axioms (such as a simple conjunction of the axioms), should he therefore come to believe every other theorem?

Another strategy is to use nonstandard modal logic. In addition to possible worlds where the logical laws hold, models of such logics can include nonclassical (impossible) worlds, where the truth value of every formula is determined separately, without any connection to the values of its subformulas. This kind of models was proposed by Rantala (1975).

Still another approach is to supplement the standard modal theory of implicit belief with a mechanism that selects some syntactic structures as believed explicitly. Each world is associated (via a function $\mathcal{A}_i(w)$ for each agent $i$) with a set of formulas that the agent is aware of in that world. The agent explicitly believes a formula $\phi$ at world $w$ iff he believes $\phi$ implicitly and is aware of $\phi$ at $w$:

$$\begin{array}{lll} (M,w) \vDash A_i\phi & \text{iff} & \phi \in \mathcal{A}_i(w) \\ (M,w) \vDash X_i\phi & \text{iff} & (M,w) \vDash K_i\phi \text{ and } (M,w) \vDash A_i\phi \end{array} \tag{4.1}$$

This approach is taken by Fagin and Halpern (1988) in their logic of general awareness.

The impossible worlds approach and the logic of awareness turn out to be equivalent in their expressive power (Wansing, 1990; Fagin et al., 2003). In both cases, one can place restrictions on the set of admissible structures to ensure that certain regularities in belief ascriptions are met. For example, in the logic of awareness, in order to make beliefs closed under conjunction simplification, one can require that the awareness function respect the condition

$$\text{if } \phi \wedge \psi \in \mathcal{A}_i(w), \text{ then } \phi \in \mathcal{A}_i(w) \text{ and } \psi \in \mathcal{A}_i(w) \tag{4.2}$$

In both cases, ‘purely semantic’ interpretation rules using possible worlds and accessibility relations are augmented with ones that directly refer to the syntax of formulas. In both cases there are no resources for the additional restrictions to make a distinction between formulas that are explicitly stored in an agent’s mind and those that are merely derivable. Therefore awareness sets/true formulas in an impossible world will inevitably be deductively closed under a limited inference system corresponding to the restrictions (for example, if (4.2) is respected, awareness sets will be closed under conjunction simplification). As we will see (p. 96), this is not necessarily a realistic assumption. On the other hand, in contrast to the approach based on the type of logic introduced in Chapter 2, logic of awareness and impossible world structures have no difficulty handling introspective beliefs.

Yet another way to make use of syntactic information is to consider structured propositions — that is, syntactic derivation trees, where every node corresponds to a constituent, and a set-theoretic meaning is assigned to it (thus, sets of possible worlds are assigned to whole sentences). The objects of belief are such structured propositions. This is the position taken in Lewis (1970); other notable proponents are Soames (1989), Moore (1995, Chapter 5). In this theory, it is impossible to distinguish synonymous leaf nodes (for example, *Cicero* from *Tully*, or *furze* from *gorse*). This leads Moore to a claim that any belief ascription involving such terms (or at least, any ascription where substitution of such terms leads to problems) is metalinguistic (p. 116–118). Otherwise, the structured propositions view is very similar to representational approaches to be discussed in the next section.

The functionalist (in Stalnaker’s sense) approach ascribes belief states on the basis of behavioural dispositions; therefore, agents with the same dispositions will have the same beliefs (and desires) ascribed to them. I will now argue that this leads to certain unwelcome results, however finely we individuate beliefs.

Dennett (1975) provided an example that purports to refute representational theories of belief. (Matthews (2007) uses it as one of his most important arguments.)

> In a recent conversation with a designer of a chess-playing program I heard the following criticism of a rival program: "It thinks it should get its queen out early." This ascribes a propositional attitude to the program in a very useful and predictive way, for as the designer went on to say, one can usually count on chasing that queen around the board. But for all the many levels of explicit representation found in that program, nowhere is anything roughly synonymous with "I should get my queen out early" explicitly tokened. The level of analysis to which the designer's remark belongs describes features of the program that are, in an entirely innocent way, emergent properties of the computational processes that have "engineering reality". I see no reason to believe that the relation between belief-talk and psychological process talk will be any more direct.

Now, there may be reasons why a program would not be disposed to answer affirmatively to sentence $D$ — 'I should get my queen out early', the most trivial being that it might not have an interface for answering questions. But take Harry, a human player who memorized the rules by which the program selects its moves. Harry would certainly respond affirmatively to any sentence that is explicitly represented by the program (in an appropriate mode), since these sentences are also represented in his mind. But there is no reason for him to affirm $D$. Moreover, it is easy to imagine that the program has a rule 'I should *not* get my queen out early', but the rule's weight in influencing the program's behaviour is too low, and in a real game it is always trumped by other rules. Still, in such a setup Harry will in fact be disposed to affirm the negation of $D$. Therefore, Dennett's example fails the test based on the disquotation principle. Suppose now that the proponent of the interpretative theory insists on ascribing the queen-should-be-out belief to the program, since it does predict all its relevant behaviour, but avoids ascribing it to Harry, since his linguistic behaviour is inconsistent with this belief. But now, let us imagine poor Harry has a stroke and loses all his linguistic abilities, retaining his chess-playing abilities. The theorist would have to say that the stroke caused him to *acquire* the queen-should-be-out belief, since now his behavioural dispositions are indistinguishable from the program's in any relevant way. I find this result highly counterintuitive.

The designer's ascription was 'useful and predictive', as Dennett puts it. This does not prevent it from being erroneous. Notice, by the way, that the designer in question ascribed beliefs to a *rival* program, not to his own creation. Dennett does not specify whether the designer had any idea what information the rival program represented explicitly. As for the usefulness of the ascription, a person who does not know enough physics will find it useful to ascribe to stones a desire to be as close to earth as possible. The ascription would still be wrong.

Another case that fails the disquotation principle is knowledge of language. A naïve speaker might not respond affirmatively to a sentence 'The subject in declarative sentences of English goes before the verb', even though he follows such a rule in his speech. This observation pushes us into denying that a speaker believes the rules of his language. Let me try to justify this. It is doubtless a part of ordinary usage to say that a person *knows* a language. But in this kind of expression the object position of the verb *to know* is occupied by a thing — a language, not by a proposition. This use of *know* cannot be analyzed as true belief, or at least not as directly as for knowledge of propositions.[4] As for statements such as 'English speakers know such and such rule of English grammar', these are obviously part of professional jargon; it takes some effort for a naïve speaker to learn such a use of *know.* It is not clear to me that such statements are strong indicators that correct use of a given rule should be considered belief in that rule. Again, one often finds cases where a native speaker believes in some rule of normative grammar (such as prohibition to split infinitives), is disposed to affirm the rule when asked, but violates it in his own speech.

### 4.1.3 Representationalism

Another approach to analyzing belief and belief attributions relies on a hypothesis that agents capable of belief operate on a sentence-like symbolic representation of propositions in their minds. The language of such representation is often called Mentalese. Perhaps the most famous defense of this view is Fodor (1975); another exposition can be found in Moore and Hendrix

[4] Some languages have two distinct verbs translated as *know*, one for knowing propositions, another for 'being acquainted' with objects and people. At least in one such language, Dutch, it is the second verb (*kennen*) that is used for knowing languages, rather than the first (*weten*). While having different translations in another language is by no means a decisive argument for ambiguity, it at least makes such ambiguity somewhat more plausible.

(1979).

The proponent of such a theory has to specify how Mentalese sentences get their truth conditions; one approach that seems satisfactory in this respect is due to Crimmins (1992) and involves the notion of normal concepts (p. 97).

In such a theory, an agent $A$ believes $p$ iff $p$'s translation into the language of thought is represented in $A$'s mind in a certain way (one talks about an agent's 'belief box' as opposed to his 'desire box', and perhaps some other boxes). The object of a belief attitude is then a sentence of the internal language (or rather, the type of such sentences).

Individuation conditions for belief under this approach coincide with individuation conditions for sentences in the internal language. Since in belief attributions these internal sentences have to be translated into sentences of external natural languages, one can explain some of the regularities in belief attributions: one internal language sentence can correspond to more than one English sentence, and both of them will be attributable to a believer. For example, one can assume that verbs *buy* and *sell* have the same internal representation, and this will explain that

(38) John believes that Michael sold a car to Peter.

entail

(39) John believes that Peter bought a car from Michael.

(This is a move available, among others, to accounts of belief based on DRT, such as Kamp (1990).)

However the number of beliefs attributable to any human agent is infinite, which makes them incapable to be represented in a finite brain, unless one adopts a highly stretched notion of what constitutes a representation. Here are some beliefs attributable to any reader of this text which are not likely to be explicitly represented in his mind (at least on the first reading):

(40) The Eiffel Tower is more than 2 meters tall.

(41) 58901 is greater than 57632.

(42) Bicycles are inedible.

This leads us to a view that I am about to defend: beliefs we ascribe to an agent are those that are derivable from his explicit beliefs using a certain restricted class of inferences. This view was first stated (and rejected, from the interpretivist positions) by Dennett (1975). A more serious defence can be found in Field (1978), but without a clear explanation of what counts as a simple inference. A formal presentation was built by Konolige (1986). His theory has a particular class of inferences attributable to an agent as a parameter. In this chapter, I attempt to fill in the value for this parameter in the natural-language *belief* predicate.

One more alternative is that of Crimmins (1992). Crimmins divorces the truth condition for *A has a belief that p*, which he considers to be true only in cases of explicit beliefs, from those of *A believes that p*. For the latter, his proposal is

> *A* believes *p* (in way $\tau$) just in case
> it is as if *A* has an explicit belief that *p*.

He considers this superior to the 'formation-dispositional' accounts (those essentially based on dispositions to assent) and 'simple consequence' accounts, such as Field's. According to Crimmins, *A believes that p* is true iff adding

$p$ to the set of $A$'s explicit beliefs would not change his cognitive dispositions (apart from some irrelevant changes such as introspection concerning explicit beliefs and micro-changes in reaction times).

One advantage of this proposal is that it at the same time makes exactly as much distinction between mental states as it is useful for predicting behaviour and does so respecting all the fine granularity of the representational account.

In my work, I am primarily interested in the construction *A believes that p*; I feel *A has the belief that p* is mostly used in technical talk (even though, of course, it did exist before this way of technical talk developed). As such, philosophers' intuitions about truth conditions of the latter construction are too easily swayed by one's preferred theory. If Crimmins only recognizes explicitly represented beliefs as objects, I am content with that.

As for the *as if* approach to tacit belief ascriptions, even if it is the right analysis, in order to make any predictions about truth conditions of belief sentences it needs to be supplemented by a theory which tells us exactly what internal sentences are such that, even though $A$ does not have them as explicit beliefs, it is *as if* he had. In other words, one still needs to know how an agent's mental state can be extended without influencing his relevant dispositions. And to do that, one has to say which explicit-belief producing operations count as "easy enough". Even though characterizing the class of allowed inferences no longer counts as conceptual analysis (which is now provided by the *as if* theory), it remains a useful endeavour. In order to capture regularities in belief ascriptions, one has to build a model for deriving explicit beliefs.

## 4.2 Inferences in belief ascription

Agents — human and otherwise — differ in their inferential capabilities. Whenever one ascribes a belief in $p$ to another agent $A$, one needs to be sure that $A$ is able to derive $p$ quickly and reliably. At a minimum, this has the consequence that inferences one is prepared to make on $A$'s behalf should be decidable.

Similarly to §3.3, in this section I aim to build as complete a list of such inferences as possible.

It seems to me that, at least with respect to human believers, the following regularities in belief ascription exist:

- [Conjunction simplification]: If $A$ believes that $p \wedge q$, then $A$ believes that $p$ and $A$ believes that $q$. This covers all the cases described in the previous chapter, namely simple conjunction, dropping veridical adverbs and dropping optional arguments. In the following examples, John's words should be interpreted as a sincere assertion.

  (43) John: La Paz is in Bolivia and Quito is in Peru.
  ∴ John believes that Quito is in Peru.

  (44) John: It is raining heavily today.
  ∴ John believes that it is raining.

  (45) John: Peter ate three slices of pizza.
  ∴ John believes that Peter ate.

- [Coreferential replacement]: If $A$ believes that $Fb$ and believes $b = c$ (where $b$ and $c$ may be either proper names or definite descriptions), then $A$ believes that $Fc$.

(46) *Background: John knows that Suzie is Bill's (only) sister.*
John: Suzie is at the movies.
∴ John believes that Bill's sister is at the movies.

One should distinguish between *de re* belief ascriptions and application of coreferential replacement to *de dicto* belief reports. Whatever truth conditions philosophical theories assign to *Lois Lane believes that Clark Kent can fly*, pre-theoretically this sentence counts as false, and said philosophical theories have to argue at length that our pre-theoretical intuitions need to be ignored in this particular case. There is no comparable difficulty when drawing the inference in (46).

The identity between $b$ and $c$ should be prominent enough for the believer; roughly, if we adopt something like Kamp's theory of propositional attitudes (Kamp, 1990), $b$ and $c$ should be part of the information stored in an internal anchor. Thus, even if John knows that the inventor of bifocals and the first Postmaster General are the same person (with neither description being for him a primary way of referring to Benjamin Franklin), Test 4.0.1 may still fail, i. e.

(47) John believes that the inventor of bifocals was a gifted engineer but he does not believe that the first Postmaster General was a gifted engineer.

may be true.

- [Existential introduction]: If $A$ believes that $Fo$, then $A$ believes that $\exists x Fx$.

(48) John: Peter is bald.
$\therefore$ John believes that there are bald people.

One should distinguish existential introduction forming a *de re* belief report describing a belief about a certain object: *There are people John believes to be bald*, from ascribed belief in an existential, as in (48).

Some restrictions will be placed on applicability of this inference (see p. 91).

- [Scalar inferences]: If $A$ believes that $o$ is at least at a point $n$ on a certain scale, and $m$ is less than $n$, then $A$ believes that $o$ is at least at a point $m$ on that scale.

(49) John: Peter is more than 2 meters tall.
$\therefore$ John believes that Peter is more than 1.80 tall.

- [Subtyping]: Inferences based on subtyping: if $A$ believes that $o$ belongs to the class $C$, and $C$ is a subclass of $D$, then $A$ believes that $o$ belongs to $D$:

(50) John: Demyan is a cat.
$\therefore$ John believes that Demyan is an animal.

I assume here that certain subclass relations are part of the ontology dictated by the conceptual system reflected in the language; everyone who possesses the subclass concept $D$ knows that it is a subclass of $C$.

The subtyping rules can be seen as a special case of the rule of universal exploitation: the statement that cats are a subtype of animals entails that every cat is an animal; after that, we derive "Demyan is an animal" from “Demyan is a cat”. However it makes sense to isolate subtyping as a separate rule, since not every instance of universal exploitation is permissible when deriving other people’s beliefs.

- [Lexical inferences]: The previous item concerns lexical elements used to express the belief — lexemes that stand in a subtype-supertype relation. There are certain other lexical inferences one is allowed to make on behalf of the believer.

  (51) John: Bill sold a car to Peter.
  $\therefore$ John believes that Peter bought a car from Bill.

  (52) John: Peter killed Tom.
  $\therefore$ John believes that Tom died.

- [‘Internal perception’]: It is unlikely that all beliefs are represented in a linguistic or language-like form. Some will be stored as maps or images, or some other way of representation. Producing an internal sentence on the basis of such representation is not inference in the usual sense, but postulating such an operation is necessary to explain some of beliefs we want to ascribe to agents. One example of this is needed in the DENY puzzle: surely every literate English speaker has a belief *The word DENY ends in E-N-Y*, but this belief is not stored as a sentence. Rather, when a need arises to make this belief explicit, one can imagine the spelling of the word and check what letters are involved; this process resembles perception of a word written on a sheet of paper.

- [Inference from ignorance]: In certain cases, absence of information to the contrary (that is, $\neg p$) is sufficient to ascribe to the agent belief that $p$.

(53) John believes that snow does not turn red when it touches the ground.

All the inferences in the foregoing list survive the absurdity test; for example, in the case of (48),

(54) John believes that Peter is bald but he does not believe that there are bald people.

feels inconsistent to me, at least if John is a human. By extension, however, beliefs are ascribed to cognitive agents other than humans, such as computer programs. The ascriber may well know that such an agent, even though its internal representational system is rich enough to have certain beliefs, lacks the power to perform the required inferences. Thus, if John is indeed a computer program that lacks the mastery of existential introduction, (54) may turn out to be true. It is a matter of lexical semantics to decide whether such use of *believe* employs the same standard sense of the word or if it is a metaphorical extension. If we decide to take the former route, regularities in belief ascription will not be part of the lexical meaning for *believe*, but rather a matter of common-sense psychology. If we choose the second option, regularities will be analytic. In my opinion, they are worth describing regardless of our decision.

## 4.3 Natural logic

Inference types listed in the previous section, as well as inferences allowed in indirect speech reports, apply to subordinated clauses as well. However in certain positions the direction of entailment is reversed. Thus, we have entailment in 'positive' positions, as in (55) and (56), and in 'negative' positions, as in (57) and (58):

(55) John believes that if it starts to rain, the road will be wet and slippery.

$\therefore$ John believes that if it starts to rain, the road will be slippery.

(56) John believes that every farmer owns a cow.

$\therefore$ John believes that every farmer owns an animal.

(57) John believes that if we buy another table, there will be too much furniture in the apartment.

$\therefore$ John believes that if we buy another table and a sofa, there will be too much furniture in the apartment.

(58) John believes that all computers are unreliable.

$\therefore$ John believes that all Macs are unreliable.

Of course, the entailments between the contents of beliefs are explained by positive or negative monotonicity of various operators. My point here is that employing such monotonicity considerations is allowed on behalf of other believing agents (at least up to a certain degree of sentence complexity).

A system (or rather a family of closely related systems) called Natural Logic is well suited to capturing exactly this kind of entailments. Natural logic, as a rule, uses syntactic trees of natural language sentences (see, for example, Zamansky et al. (2006)). Every type $\tau$ of constituents in the grammar is equipped with a partial order $\preccurlyeq_{\tau}$ ($\preccurlyeq$ should be read as "less general than"); one can derive whether two elements of a given type stand in this relation by looking at whether this relation holds between their subconstituents. For the type of sentences, $t$, the partial order $\preccurlyeq_{t}$ is interpreted as entailment.

In this work, however, I use neo-Davidsonian logical formulas as the internal representation of natural language sentences. This is done primarily out of convenience, because there is very little, if any, data on what the actual internal language of thought might look like. The language of such formulas has an extremely simple syntax and an uncomplicated type system. Thus the task of building a deduction system for this language in the style of natural logic is much simpler than for any actual natural language.

In first order logic, there are just three kinds of expressions — object-denoting (constants and, if allowed, free variables), predicates and formulas. Object-denoting expressions are not comparable in generality.[5] Predicates of the same arity can be compared (**adore** $\preccurlyeq$ **like**), and $\preccurlyeq$ for formulas is defined on the basis of $\preccurlyeq$ for predicates using the following rules:

$$\phi \preccurlyeq \phi \text{ N-}\textsc{Refl}$$

$$\frac{\phi \preccurlyeq \psi \qquad \phi \preccurlyeq \chi}{\phi \preccurlyeq \chi} \text{ N-}\textsc{Trans}$$

$$\frac{P^n \preccurlyeq Q^n}{P^n(c_1, \ldots, c_n) \preccurlyeq Q^n(c_1, \ldots, c_n)} \text{ N-}\textsc{Pred}$$

$$\phi \wedge \psi \preccurlyeq \phi \text{ N-}\textsc{And1} \qquad \phi \wedge \psi \preccurlyeq \psi \text{ N-}\textsc{And2}$$

$$\frac{\phi_1 \preccurlyeq \phi_2 \qquad \psi_1 \preccurlyeq \psi_2}{\phi_1 \wedge \psi_1 \preccurlyeq \phi_2 \wedge \psi_2} \text{ N-}\textsc{Conj}$$

$$\frac{\phi \preccurlyeq \psi}{\neg\psi \preccurlyeq \neg\phi} \text{ N-}\textsc{Neg}$$

$$\frac{\phi_2 \preccurlyeq \phi_1 \qquad \psi_1 \preccurlyeq \psi_2}{\phi_1 \rightarrow \psi_1 \preccurlyeq \phi_1 \rightarrow \psi_2} \text{ N-}\textsc{Cond}$$

$$d_1 = d_2 \preccurlyeq d_1 \leq d_2 \text{ N-}\textsc{Eq}$$

[5] I assume that degrees are objects, and that degrees with the same unit of measurement can be compared by $\leq$. I will have rules transforming $\leq$ comparisons into $\preccurlyeq$ comparisons as formulas.

$$\frac{d_1 \leq d_2}{c \leq d_1 \preccurlyeq c \leq d_2}\ \text{N-Neq1} \qquad \frac{d_1 \leq d_2}{d_2 \leq c \preccurlyeq d_1 \leq c}\ \text{N-Neq2}$$

$$\frac{\phi(c) \preccurlyeq \psi(c)}{\exists x\phi(x) \preccurlyeq \exists x\psi(x)}\ \text{N-Exist} \qquad \frac{\phi(c) \preccurlyeq \psi(c)}{\forall x\phi(x) \preccurlyeq \forall x\psi(x)}\ \text{N-ForAll}$$

Some of these rules deserve a short comment. N-And1 and N-And2 take care of everything that translates as conjunction into our representation, including intersective adjectives, relative clauses and argument dropping. Unfortunately, subsective adjectives are not representable in first-order language, so this logic cannot deal with them. Once we extend the language, we would need an extension of the logic; the system would then perhaps be closer to that of (Zamansky et al., 2006), with its types annotated for monotonicity of functions, rather than to this simple sketch.

As for degrees, I assume that expressions like 'John is less than two meters tall' translate as $\exists d(\mathbf{tall}(\mathbf{j}, d) \wedge d \leq \mathbf{2m})$. That is, $d$ is the *exact* degree to which John is tall. There is an unlimited supply of degree constants for each unit of measure (natural numbers, for counting quantities of discrete objects, are just another unit of measure). Using degrees, we can take care of all the scalar inferences.

Let us see how specificity relations between sentences (represented as formulas) can be derived. A derivation of 'Every student whom Mary touched smiled' $\preccurlyeq$ 'Every student whom Mary kissed smiled,' that is,

$$\begin{gathered}\forall x((\mathbf{student}(x) \wedge \mathbf{touch}(\mathbf{m}, x)) \rightarrow \mathbf{smile}(x)) \\ \preccurlyeq \\ \forall x((\mathbf{student}(x) \wedge \mathbf{kiss}(\mathbf{m}, x)) \rightarrow \mathbf{smile}(x))\end{gathered}$$

is presented on Figure 4.1.1 (compare Zamansky et al. (2006, p. 285)) (Here I represent multi-argument predicates in their standard form, without splitting

1. ‘Every student whom Mary touched smiled’ $\preccurlyeq$ ‘Every student whom Mary kissed smiled,’

$$\dfrac{\dfrac{\begin{array}{c}\mathbf{student}(c) \preccurlyeq \\ \mathbf{student}(c)\end{array}\ \textsc{N-Refl} \qquad \dfrac{\mathbf{kiss} \preccurlyeq \mathbf{touch}}{\begin{array}{c}\mathbf{kiss}(\mathbf{m}, c) \preccurlyeq \\ \mathbf{touch}(\mathbf{m}, c)\end{array}}\ \textsc{N-Pred}}{\mathbf{student}(c) \wedge \mathbf{kiss}(\mathbf{m}, c) \ \preccurlyeq\ \mathbf{student}(c) \wedge \mathbf{touch}(\mathbf{m}, c)}\ \textsc{N-Conj} \qquad \begin{array}{c}\mathbf{smile}(c) \preccurlyeq \\ \mathbf{smile}(c)\end{array}\ \textsc{N-Refl}}{\dfrac{\begin{array}{c}(\mathbf{student}(c) \wedge \mathbf{touch}(\mathbf{m}, c)) \to \mathbf{smile}(c) \ \preccurlyeq \\ (\mathbf{student}(c) \wedge \mathbf{kiss}(\mathbf{m}, c)) \to \mathbf{smile}(c)\end{array}}{\forall x((\mathbf{student}(x) \wedge \mathbf{touch}(\mathbf{m}, x)) \to \mathbf{smile}(x)) \ \preccurlyeq\ \forall x((\mathbf{student}(x) \wedge \mathbf{kiss}(\mathbf{m}, x)) \to \mathbf{smile}(x))}\ \textsc{N-ForAll}}\ \textsc{N-Cond}$$

2. ‘John is exactly 1.7m tall’ $\preccurlyeq$ ‘John is less than 2m tall’

$$\dfrac{\dfrac{\dfrac{d = \mathbf{1.7m} \preccurlyeq d \leq \mathbf{1.7m}\ \textsc{N-Eq} \qquad \dfrac{\mathbf{1.7m} \leq \mathbf{2m}}{d \leq \mathbf{1.7m} \preccurlyeq d \leq \mathbf{2m}}\ \textsc{N-Neq1}}{d = \mathbf{1.7m} \preccurlyeq d \leq \mathbf{2m}}\ \textsc{N-Trans}}{\mathbf{tall}(\mathbf{j}, d) \wedge d = \mathbf{1.7m} \ \preccurlyeq\ \mathbf{tall}(\mathbf{j}, d) \wedge d \leq \mathbf{2m}}\ \textsc{N-Conj}}{\exists y(\mathbf{tall}(\mathbf{j}, y) \wedge y = \mathbf{1.7m}) \ \preccurlyeq\ \exists y(\mathbf{tall}(\mathbf{j}, y) \wedge y \leq \mathbf{2m})}\ \textsc{N-Exist}$$

Figure 4.1: Derivations in Natural Logic

them in a neo-Davidsonian fashion, in order to increase readability.) Another example, involving degrees, is presented on Figure 4.1.2.

Partial order between elementary expressions of the language represents relations between concepts; these statements are analytic truths accessible to all agents.

Lexical inferences other than subtyping can be added as meaning postulates:

$$\mathbf{buy}(c_1, c_2, c_3) \preccurlyeq \mathbf{sell}(c_3, c_2, c_1)$$

$$\mathbf{kill}(c_1, c_2) \preccurlyeq \mathbf{die}(c_2)$$

There is a certain tension between this kind of postulates and neo-Davidsonianism: in neo-Davidsonian presentation the rule for *buy* and *sell* will look something like

$$\begin{array}{c} \mathbf{buy}(e_1) \wedge \mathrm{Agent}(e_1, c_1) \wedge \mathrm{Theme}(e_1, c_2) \wedge \mathrm{Seller}(e_1, c_3) \\ \preccurlyeq \\ \mathbf{sell}(e_2) \wedge \mathrm{Agent}(e_2, c_3) \wedge \mathrm{Theme}(e_2, c_2) \wedge \mathrm{Buyer}(e_2, c_1) \end{array}$$

We need to be able to recognize possible reorderings among the conjuncts, so an additional rule is required:

$$\phi \wedge \psi \preccurlyeq \psi \wedge \phi$$

In order to incorporate our brand of natural logic into the dynamic logic of inferences, we stipulate that an elementary action allows an agent to move from an explicit belief in a more specific sentence to an explicit belief in a more general one:

$$\frac{\mathbf{B}_a\phi \qquad \phi \preccurlyeq \psi}{\langle \mathrm{N}_a \rangle \mathbf{B}_a\psi}\ \mathrm{N}$$

It should be stressed that, apart from being a simplified toy, my system differs from other presentations of Natural Logic in one important respect. Typically proponents of Natural Logic aim for as broad a class of inferences as possible; the power of the logic is only restricted by what patterns of deduction are representable in the surface syntax of natural languages. My purpose is to capture exactly those inferences that are used in deriving belief ascriptions, i. e., a very limited class. For example, I don't want to include the rule for 'every' from (Zamansky et al., 2006) (which allows them to perform derivations like 'Mary kissed every student', 'No student whom Mary kissed walked' ⊢ 'No student walked'), since it would take my system beyond the limits of belief ascription.

## 4.4 Indirect speech, revisited

In Chapter 3, I stated that indirect speech supports dropping focus markers and conjunction simplification as operations on its content. However almost every conjunction, in particular when using a neo-Davidsonian internal language, will be in scope of some quantifiers. In order to allow such an operation, I need to resort to the same kind of Natural Logic as for belief, with two restrictions:

1. Rules are applied only in monotone increasing contexts; this means N-Neg is dropped and N-Cond has a simpler form:

$$\frac{\psi_1 \preccurlyeq \psi_2}{\phi \to \psi_1 \preccurlyeq \phi \to \psi_2}\ \textsc{N-Cond'}$$

2. The lexical material used in indirect ascription should be a subset of the material used in the primary utterance (after translation between

languages, if any).

Both conditions are necessary, as witnessed by the inadmissibility of the following examples:

(59) John: If Mary or Kate come, I will be happy.
Peter: $^?$John says that if Kate comes, he will be happy.

(60) John: Bill saw a dog.
Peter: $^\#$John says that Bill saw a dog or a cat.

In (59), restriction 1 is violated, but not 2. In (60), it's the other way around.

This new analysis allows us to take care of the focus operators mentioned on p. 59. Take the case of *only.* Its meaning (ignoring the distinction between presupposition and assertion) can be expressed as

$$[\![\textit{only a did P}]\!] = Pa \wedge \forall x(x \neq a \rightarrow \neg Px)$$

We can simplify this conjunction by dropping either its first or second conjunct. If the first conjunct is retained (which corresponds to the presupposition of the primary utterance), it can be expressed using material from the original sentence and thus respecting condition 2.

One problem is that if we retain the second conjunct, we cannot express it using only words from the primary utterance, thus condition 2 is violated. But the second conjunct *can* be reported in indirect speech:

(61) John says that nobody but Michael solved the problem.

(I thank both Nick Asher and Martina Faller for independently overriding my theory-laden intuitions.) This is not predicted by my current account.

## 4.5 Problems

We now return to the case of belief ascriptions.

Natural Logic allows us to ascribe certain beliefs to an agent $A$ on the basis of some evidence about his explicit beliefs. However not all such ascriptions will be predicted by our brand of natural logic, and not all the ascriptions predicted by this logic are intuitively justified. We consider these problematic cases.

First, even if individual steps in a natural logic-based inference are unproblematic, once it becomes sufficiently long, the plausibility of belief ascription diminishes. Consider[6]

(62) John: Every girl has a small dog.
$\therefore$ John believes that every small girl has a dog.

This inference can be obtained by applying Natural Logic rules twice: first to get

$$\text{every girl has a small dog} \preccurlyeq \text{every girl has a dog}$$

and then

$$\text{every girl has a dog} \preccurlyeq \text{every small girl has a dog}$$

For each of these steps, the belief ascription of the right sentence to someone who explicitly believes the left one seems fully justified. But the inference in (62) is already somewhat dubious. It is not hard to produce more elaborate examples, further decreasing acceptability of the inference.

[6] This is based on an example from Lauri Karttunen's talk at Texas Linguistic Society conference 2009.

One could appeal to the distinction between competence and performance here. But whereas this distinction provides a way to explain why real speakers' grammatical judgments differ from the grammaticality prediction of a theory that concerns idealized speakers, in our case bringing up the notion of an idealized believer seems misguided. After all, we already have one such notion — a logically omniscient agent of the standard epistemic logic. It makes no sense to invent a much more complicated but still idealized construction.

Perhaps the way out is to restrict the number of applications of our natural logic rules to two or three iterations. Luckily, our dynamic belief logic is powerful enough for that.

Another problematic set of examples where our natural logic rules are overgenerating concerns existential introduction. The DENY problem by Powers (see p. 65) is a case in question. Thus, while (63) is intuitively a valid inference, (64) is not, even though they have the same structure.

(63) John believes that his best friend Peter is bald.
$\therefore$ John believes that there are bald people.

(64) John believes that DENY is a four-letter English word ending in E, N, Y.
$\not\therefore$ John believes that there is a four-letter word in English ending in E, N, Y.

David Beaver (p. c.) suggested that a belief ascription of $p$ to $A$ is valid when there is a computationally cheap algorithm for computing the truth value of $p$ given the explicit beliefs $A$ has[7]. Of course, this idea needs a more detailed

[7] David points out that I misconstructed his proposal. One should only count algorithms

specification of how the computational cost of an algorithm is evaluated and what counts as cheap (this will inevitably be a vague notion). However in any case it seems that the DENY case can serve as a counterexample to this proposal. After all, one only needs to run a test over 26 instances to acquire the belief in question, while the number of John's non-bald friends can easily be much larger; thus the computational cost in (63) can be higher than in (64).

A more promising account of this problem is based on the fact that in the DENY case the belief that serves as the base for existential generalization is not explicit; rather, it is of the kind obtained by 'internal perception' (John knows how the word DENY is spelled; he can imagine it written and count the letters). As far as I can see, all puzzles of this sort involve this kind of translation into linguistic internal representation from a non-linguistic medium. 'Belief box', the mechanism storing explicit beliefs, undoubtedly possesses some internal structure beyond a simple list of sentences assumed by our simplistic model. This structure likely includes efficient indexing mechanisms that allow an agent to find the needed belief quickly. Non-linguistic representations may not have such indexing; thus it is much harder to access just the piece of information we need in order to apply existential generalization (after transforming it into an internal-language sentence).[8]

One consequence of such a solution is that the inference action that will figure in our definition of belief will not be closed under repetition — rather,

that we, humans, are wired to compute easily. In this case, I consider it my task to investigate which particular algorithms for easy computation of beliefs we possess.

[8] The idea that 'internal perception' is involved in all puzzles is the result of a discussion with Joey Frazee.

it will be a disjunction of internal perception and the closure of everything else (natural logic and inference from ignorance).

Next, we have inferences 'from ignorance'.[9] Some of these may be handled as subtyping inferences. For example,

(65) John believes that bicycles are inedible.

can be derived using **bicycle** $\preccurlyeq$ **mechanism** and *John believes that mechanisms are inedible*, which, in turn, one can plausibly take to describe an explicit belief John has. However others remain problematic. As a first approximation, the inference rule employed in deriving such beliefs seems to be

$$\frac{\neg\mathbf{B}p}{\mathbf{B}\neg p}\ \textsc{Ign}$$

It is not clear, however, for which $p$ this rule should be applicable. Clearly not all $p$-s will work (otherwise the believer would operate under a delusion of his own omniscience). Moreover, it is clearly impossible to determine which beliefs belong to the requisite class on the basis just of their linguistic properties –

[9] Inferences from ignorance are used as an argument against the deductive theory of belief in (Lycan, 1985; Crimmins, 1992).

Moore (1985) discusses this type of problematic inferences, using them as justification for his autoepistemic logic:

> Consider my reason for believing that I do not have an older brother. It is surely not that one of my parents once casually remarked, "You know, you don't have any older brothers," nor have I pieced it together by carefully sifting other evidence. I simply believe that if I did have an older brother I would know about it; therefore, since I don't know of any older brothers, I must not have any.

things like cultural background should be taken into account. For example, if Misha is a Russian with basic school-level education, one is justified in stating

(66) Misha believes that Hungarians never conquered Moscow.

At the same time, one would not be justified in ascribing the same belief to Michael, who lives in America and does not count Russian history among his areas of interest.

Thus, the first amendment to the rule will be

$$\frac{\neg\mathbf{B}p \qquad Kp}{\mathbf{B}\neg p}\,\textsc{Ign}$$

where the interpretation of $K$ is roughly '$A$ believes he would have known $p$ if it were true'. The 'believes' here should be different from the explicit belief operator $\mathbf{B}$ to avoid vicious circularity. $K$ is a meta-level operator; there should be a computationally cheap procedure which lets $A$ determine whether $K$ applies to a given sentence $p$.

A further, less disruptive amendment is needed because 'would have known' should include tacit beliefs. I have an explicit belief that the Ostankino tower is 540 meters tall; my belief that it is more than 10 meters tall is tacit, derivable by a scalar inference. However I should not be able to use the Ign rule to produce a belief that the tower is not more than 10 meters tall. The rule now becomes

$$\frac{\neg\langle\textsc{Bel}\backslash\textsc{Ign}\rangle\mathbf{B}p \qquad Kp}{\mathbf{B}\neg p}\,\textsc{Ign}$$

where Bel denotes the action associated with belief ascriptions (which will include Ign in its definition), and Bel\Ign denote this action with the mention of Ign removed.

## 4.6 Examples

Now that my position with respect to belief ascriptions is mostly stated, I would like to run through a list of examples that were considered problematic by various researchers, and indicate how my theory handles them. Most of the examples have already been mentioned, but I would like to have the whole collection in one place to make evaluating my position easier.

- The opinionated man (Lycan, 1985). This is someone who responds positively or negatively to questions depending on something totally irrelevant to the question's content (such as weather or random number generator results). The example is meant to serve as a refutation of the Carnap-style dispositional theory of belief. Since our theory does not predict belief based on disposition to assent (even though it does treat lack of such disposition as evidence against presence of belief), the example poses no threat to us.

- The excited speaker (Audi, 1982). A person speaks at a dinner table. If someone asks him whether he talks too loudly, he would immediately realize that he does and assent. Still, it is counterintuitive to ascribe to the excited speaker the belief *I am talking too loudly.* Again, this is primarily a counterexample to the dispositional theory. Realization that the voice is too loud involves perception, which in our theory is not an operation allowed as a step in making a tacit belief explicit.

- Quick and slow thinker (Lycan, 1985). This is an argument against a variant of the deduction theory that allows a belief ascription of $p$ to $A$ if $A$ is able to infer $p$ quickly enough (Lycan sees this as one of the ways to make the theory of Field (1978) more precise). One would then ascribe more beliefs to an agent who is able to think faster, which, according to Lycan, is counterintuitive. Our theory does not measure ease of inference in terms of time, so this example poses no threat.

- Easy inference (Lycan, 1985). Another precisification Lycan proposes for Field's theory is to measure ease of inference as the number of proof steps. He rightly observes that this makes the notion of belief dependent on a particular proof system. This objection is partially applicable to my theory. However, since it is not the number of rule applications that counts, but patterns determining proof shape, these patterns may well be intertranslatable between different proof systems. Moreover, the fact that these particular patterns are able to predict the applicability of belief ascriptions serves as evidence that the proof system used by actual speakers at least resembles the one I employ.

- Chess program (Dennett, 1975; Matthews, 2007). See p. 71. In short, I deny that belief ascription is true in this case.

- Puzzles (Powers, 1978). See p. 91 for the DENY example. It seems that all the problematic cases involve internal perception followed by an application of the existential introduction rule. I forbid this pattern of inferences in my formal analysis. It should be noted, however, that as a result my system is no longer closed under deduction, and ceases to be a special case of the approach described in Konolige (1986).

- Absent-mindedness (Crimmins, 1992). $A$, who has put his key in his pocket five minutes ago is now searching for it. Does he believe that the key is in his pocket? It seems that both answers "yes" and "no" have some support; on the one hand, the sentence *The key is in my pocket* is likely to be represented somewhere in $A$'s mind; on the other hand, it does not seem to be accessible in a way that allows him to use it in guiding his actions, including perhaps assent or dissent.

My theory does not have anything to say here that would distinguish it from any other kind of representationalism. Most likely, the token of the sentence in question that is represented in $A$'s mind does not occur in the "belief box" — even though the way of representation is the same as for beliefs rather than, say, desires, the contents of the belief box have to be accessible to the believer. Note also that the disquotational principle does not work here.

- Elder brother (Moore, 1985; Lycan, 1985). See the discussion on p. 93. This kind of tacit beliefs represents a genuine challenge, and even though I tried to capture the kind of reasoning needed in deriving them, I am not satisfied with the results.

- Eaten bicycle (Crimmins, 1992). Examples like *I have never eaten a bicycle* are similar to the elder brother cases (p. 93), but not as hard, since there is at least hope to reduce them to sybtyping.

## 4.7 Formal presentation

The formal representation of truth conditions for *de dicto* belief ascriptions is very similar to that of indirect speech. The difference is just the type of inferences allowed. Similarly to indirect speech, quantification over internal-language sentences is needed.

$$[\![ A \textit{ believes that } q ]\!] = \exists p(B_A p \wedge \Box(\mathbf{B}p \rightarrow \langle\ \mathrm{IP} \cup Cl(\mathrm{N}, \mathrm{CR}, \textsc{Ign})\rangle \mathbf{B}q))$$

(N for natural logic derivations, CR for coreferential replacement, Ign for inferences from ignorance, IP for 'internal perception'.)

## 4.8 Concluding remarks

In this chapter, I investigated the range of inferences that are supported in belief ascriptions. These are derivations that the ascriber can perform on behalf of the believer. The derivations should be ones that are easy for humans to perform. It should also be easy for human agents to find an inference with a given result. To repeat, these are inferences such that for each of them, an addition of the conclusion to the set of an agent's explicit beliefs will not change his behaviour (Crimmins, 1992).

The class of allowed inferences seems rather heterogenic. Together they provide an important window at the way information is represented in human minds and at the ways it is used. One can choose among various proposed theories of information representation depending on how easy the proposed representation makes it to perform inferences that are in fact easy for people. For example, representations based on DRT with internal anchors (Asher, 1986; Kamp, 1990) make it very easy to compute conjunction simplification, existential introduction and replacement of coreferential NPs, but not scalar inferences.

It is instructive to compare the range of inferences for belief reports and those for indirect speech reports.

| | **Indirect speech** | **Belief reports** |
|---|---|---|
| Conjunction simplification | yes | yes |
| Dropping subsective adjectives | yes | yes |
| Dropping veridical adverbs | yes | yes |
| Dropping focus-sensitive operators | yes | yes |
| Existential generalization | no$^*$ | yes |
| Scalar inferences | no$^*$ | yes |
| Subtyping | no$^*$ | yes |
| Other lexical inferences | no$^*$ | yes |
| NL inferences in decreasing contexts | no | yes |
| Coreferential replacement | masked by *de re* readings (see p. 46) | yes |
| "Internal perception" | no | yes |
| Inference from ignorance | no | yes |

Inference classes marked with $^*$ are among those which, I think, fail Test 3.1.1 for indirect speech, but which are sometimes used in linguistic practice, and are mentioned as valid at least in some of (Soames, 1989; Cappelen and Lepore, 1997; Brasoveanu and Farkas, 2007). Those without the asterisk are definitely out, whichever way one chooses of precisifying the vague speaker intuitions.

Other propositional attitude verbs differ in the range of supported inferences, and each of them should be studied separately. For example *wish* arguably supports scalar inferences:

(67) John wishes to win more than a million in a lottery.
$\therefore$ John wishes to win more than fifty thousand in a lottery.

but not conjunction simplification:

(68) Peter wishes that he wins the lottery and that the prize be more than

a million.

$\not\therefore$ Peter wishes that the prize is more than a million.

# Chapter 5

# Evidentiality in Bulgarian

I start this chapter by introducing the topic of evidentiality (Section 5.1) and explaining why it is one of the constructions that fit the general subject of this work. After that, two brief surveys follow: first, I provide a sketch of the Bulgarian verb system and the place of evidential markers in it (Section 5.2); second, I outline the current theoretical approaches to the contribution of evidentials to the meaning of a sentence (Section 5.3). Finally, after such an extensive introduction, I show how the theoretical apparatus of this thesis can help enlighten what the meaning of evidential markers really is (Section 5.4).

## 5.1 Evidentiality

Evidentiality is the linguistic encoding of the source of information. In languages such as English, evidential meanings are expressed by lexical means — adverbs like *reportedly*, *allegedly*; some uses of the epistemic *must* (von Fintel and Gillies, 2009a). In many other languages there are grammatical morphemes expressing evidentiality. In some languages evidentiality marking is obligatory on every independent clause; in others, such markers are optional. Studies of evidentiality from the point of view of linguistic typology include (Chafe and Nichols, 1986; Aikhenvald, 2004; Hrakovskiy, 2007).

Among meanings expressed by evidential morphemes we find visual evidence, auditory evidence, other sensory evidence, performative (knowledge based on speaker's own plans), internal perception, quotative (marking of indirect speech), hearsay, folklore, inference from results and inference from reasoning (Willett, 1988; Aikhenvald, 2004). Of course, in each particular language some of these possible sources are grouped together. The richest systems with obligatory evidentiality coding have up to 5 distinct grammemes, e. g. Tariana (Aikhenvald, 2004, p. 1–3), Tuyuca (Barnes, 1984), Kashaya (Oswalt, 1986)[1].

Evidentiality is often an areal phenomenon, with languages that obligatorily express it forming contiguous regions, despite having distinct genetic origins (Aikhenvald, 2004, p. 288–299). One such region spans from the Balkans in the West through Central Asia to Siberia in the East. In this area evidential markers tend to be historically derived from perfect morphemes. Another large region where rich evidential systems are widespread is South America. Evidentials appear to be a language feature that is relatively easy to borrow from one language to another.

As the title of (Chafe and Nichols, 1986) '*Evidentiality: the linguistic coding of epistemology*' clearly suggests, evidentiality is one of the most interesting grammatical categories from the philosophical point of view. One can find striking parallels between sources of knowledge recognized by philosophers, such as in the Indian Nyaya tradition, and grammatical markers in certain languages, such as the system of Cuzco Quechua (Faller, 2002) —

[1] For Kashaya, one could perhaps argue that two types of Inferential should be distinguished and/or that Visual should be counted separately from Factual, bringing the number of grammemes up to 7.

observation, hearsay, inference. Even if language grammars cannot serve as reliable indications what the structure of knowledge really is, they show us what the worldviews embodied in these languages take it to be. Similarly to natural language metaphysics (Bach, 1986) one can meaningfully speak of natural language epistemology, and systems of evidentials provide the clearest window into its architecture.

The reason evidentials are interesting for my purposes is that many of them fail to be closed under logical consequence. If a language has both a direct observation evidential marker and a marker for inference from results, then, upon seeing an empty bottle in John's room, one is typically justified to utter *There is an empty bottle in John's room-DIR*, and also *John drank some wine-INFER*. While the second of these utterances is made on the basis of the same information as the first, direct evidential marker is not appropriate.[2] Therefore a question arises what kind of inferences the speaker is allowed to perform while still maintaining the direct evidential marker, and which inferences require a switch to the inferential evidential.

The contrast between direct and inferential evidentials is, however, not so easy to investigate. It is hard, if at all possible, to determine what sentences in the speaker's language of thought encode his perceptual experience, and whether that encoding is linguistic at all.[3] On the other hand, when the

[2] One has to exclude the possibility that the evidential marker is used simply as an epistemic modal.

[3] But see (Tatevosov, 2007, p. 374–379) for an enlightening discussion. Tatevosov argues that 1.) inference from results (in Bagvalal) is not limited to lexically encoded process-result regularities; and 2.) amount of background knowledge the speaker possesses can influence whether he can draw the inference that underlies the use of the evidential.

Also see §5.3.3 for the discussion of Koev (2010), where a convincing argument is made that the main meaning contribution of the Bulgarian direct evidential (Indicative) does not concern "direct perceptual experience", but rather the time when the speaker learned about inference in question is performed on the basis of *hearsay* information, its premises are explicitly provided to us in the form of the original utterance(s), so we can ask the same question regarding the contrast between hearsay evidential markers and other means of conveying information — inference evidentials or epistemic modals.

So our question becomes: how far can one diverge from the words of a report while still attaching the hearsay marker to the information derived from that report?

## 5.2 Bulgarian data

Before taking on the topic of Bulgarian evidentiality, I would like to make a short introduction into the system of Bulgarian verb forms. My principal sources were Andrejčin (1944),[4] Scatton (1984). A comprehensive description of the Bulgarian evidential system can be found in Nitzolova (2007) (in Russian). Kutzarov (1994) has an excellent historical overview. Native speaker judgments in this chapter were provided by Ivan Derzhanski, Bojan Popov, Andrejka Lechev, Boris Doynov and Iordan Ganev. I would like to express my sincere gratitude to my Bulgarian consultants.

Bulgarian has a rich system of tenses; its verbal system is more complex than that of any other Slavic language. There are nine tenses: Present, Aorist, Imperfect, Perfect, Pluperfect, Future, Future Perfect, Past Future and Past Future Perfect. Present, Aorist and Imperfect are formed synthetically, each with its own set of person/number endings. Perfect tenses are formed by

the event in question.

[4] I used the Russian translation Andrejčin (1949), since it is easier for me to read Russian than Bulgarian.

combining an auxiliary *săm* ('be') with a participle; Future tenses are formed by combining another auxiliary *šta* (historically derived form a verb meaning 'want') with a finite form of the main verb (in some of the future tenses, together with a particle *da*). Negative forms of Future tenses employ negative forms of the verb *imam* 'have'. For a full list of indicative forms, see Tables 5.1 and 5.2. Bulgarian also has Imperative and Conditional moods.

### 5.2.1 Types of evidentials. Morphology

There are three classes of forms in Bulgarian that have been classified as indirect evidentials: hearsay, inferential, and hearsay with negative attitude[5] expressed by the actual speaker. The hearsay evidential is called "preiskazno naklonenie" in the Bulgarian tradition; Scatton (1984) translates this as Renarrated mood. For the inferential[6], Demina (1970) has coined the name Conclusive, and this term is adopted by Nitzolova (2006, 2007). Nitzolova also uses the term Dubitative for hearsay with negative attitude. Indicative is used as the name of the forms that mark direct evidence (in the past tenses) or are neutral with respect to the type of evidence (in the present and future tenses). This is the terminology that I will employ in this chapter.[7]

Indirect evidential forms only distinguish five tenses instead of nine in the Indicative: forms with reference point in the present are not distinguished from forms with reference point in the past. Thus, evidential verb forms for Imperfect coincide with those for Present; forms for Pluperfect with Perfect,

[5] Normally disbelief, but sometimes also expression of inappropriateness.

[6] As we shall see, this form does not actually always signal inference.

[7] In example translations, I will use the parenthetical *(reportedly)* to indicate Renarrated, and *(apparently)* or epistemic *must* for Conclusive. Verbs in Indicative will not be marked. The reader should keep in mind that these translations are extremely approximate.

Past Future with Future, and Past Future Perfect with Future Perfect.[8] Aorist forms are not ambiguous (within the evidential subsystem); they are also the most frequent ones in texts.

Evidential forms in Bulgarian are historically derived from Perfect forms; all of them contain an *l*-participle, either of the main verb, or of one of the auxiliaries.

Renarrated is formed by replacing the head of the corresponding indicative form (verb or auxiliary) with its *l*-participle and adding the present tense of the auxiliary *săm* 'be' in the 1st and the 2nd person. There is no auxiliary in the 3rd person. Conclusive is exactly the same, except the auxiliary in the 3rd person is retained. Dubitative is formed by applying the Renarrated-forming operation to the *săm* auxiliary of the Conclusive. Thus, Dubitative looks formally as Renarrated of the Conclusive, and this is the way it is analyzed, for example, by Kutzarov (1994). The semantics of Dubitative, however, is most certainly not derived compositionally from the semantics of Renarrated and Conclusive.

According to most descriptions, Conclusive is only used in past tenses. Maslov (1981) does provide several examples for Present and Future, but these examples do not sound natural to my informant. Therefore, these forms are absent from Tables 5.1 and 5.2, borrowed from Nitzolova (2006), p. 40–41.

[8] It is not clear whether some of these forms exist in the Conclusive; see below.

| Indicative | Conclusive | Renarrated | Dubitative |
|---|---|---|---|
| *Present* | | | |
| 1. piša | | pišel săm | pišel săm **bil** |
| 2. pišeš | | pišel si | pišel si **bil** |
| 3. piše | | pišel | pišel **bil** |
| 1. pišem | | pišeli sme | pišeli sme **bili** |
| 2. pišete | | pišeli ste | pišeli ste **bili** |
| 3. pišat | | pišeli | pišeli **bili** |
| *Imperfect* | *Imperfect = Present* | | |
| 1. pišex | pišel săm | pišel săm | pišel săm **bil** |
| 2. pišeše | pišel si | pišel si | pišel si **bil** |
| 3. pišeše | pišel **e** | pišel | pišel **bil** |
| 1. pišexme | pišeli sme | pišeli sme | pišeli sme **bili** |
| 2. pišexte | pišeli ste | pišeli ste | pišeli ste **bili** |
| 3. pišexa | pišeli **sa** | pišeli | pišeli **bili** |
| *Future* | | | |
| 1. šte piša | | štjal săm da piša | štjal săm **bil** da piša |
| 2. šte pišeš | | štjal si da pišeš | štjal si **bil** da pišeš |
| 3. šte piše | | štjal da piše | štjal **bil** da piše |
| 1. šte pišem | | šteli sme da pišem | šteli sme **bili** da pišem |
| 2. šte pišete | | šteli ste da pišete | šteli ste **bili** da pišete |
| 3. šte pišat | | šteli da pišat | šteli **bili** da pišat |
| *Past Future* | *Past Future = Future* | | |
| 1. štjax da piša | štjal săm da piša | štjal săm da piša | štjal săm **bil** da piša |
| 2. šteše da pišeš | štjal si da pišeš | štjal si da pišeš | štjal si **bil** da pišeš |
| 3. šteše da piše | štjal **e** da piše | štjal da piše | štjal **bil** da piše |
| 1. štjaxme da pišem | šteli sme da pišem | šteli sme da pišem | šteli sme **bili** da pišem |
| 2. štjaxte da pišete | šteli ste da pišete | šteli ste da pišete | šteli ste **bili** da pišete |
| 3. štjaxa da pišat | šteli **sa** da pišat | šteli da pišat | šteli **bili** da pišat |
| *Perfect* | | | |
| 1. pisal săm | bil săm pisal | bil săm pisal | |
| 2. pisal si | bil si pisal | bil si pisal | |
| 3. pisal e | bil **e** pisal | bil pisal | |
| 1. pisali sme | bili sme pisali | bili sme pisali | |
| 2. pisali ste | bili ste pisali | bili ste pisali | |
| 3. pisali sa | bili **sa** pisali | bili pisali | |

Figure 5.1: Forms of the verb *piša* ‘to write’, part 1

| Indicative | Conclusive | Renarrated | Dubitative |
|---|---|---|---|
| *Pluperfect* | *Pluperfect = Perfect* | | |
| 1. bjax pisal | bil săm pisal | bil săm pisal | |
| 2. beše pisal | bil si pisal | bil si pisal | |
| 3. beše pisal | bil **e** pisal | bil pisal | |
| 1. bjaxme pisali | bili sme pisali | bili sme pisali | |
| 2. bjaxte pisali | bili ste pisali | bili ste pisali | |
| 3. bjaxa pisali | bili **sa** pisali | bili pisali | |
| *Future Perfect* | | | |
| 1. šte săm pisal | | štjal săm da săm pisal | štjal săm **bil** da săm pisal |
| 2. šte si pisal | | štjal si da si pisal | štjal si **bil** da si pisal |
| 3. šte e pisal | | štjal da e pisal | štjal **bil** da e pisal |
| 1. šte sme pisali | | šteli sme da sme pisali | šteli sme **bili** da sme pisali |
| 2. šte ste pisali | | šteli ste da ste pisali | šteli ste **bili** da ste pisali |
| 3. šte sa pisali | | šteli da sa pisali | šteli **bili** da sa pisali |
| *Past Future Perfect* | *Past Future Perfect = Past Perfect* | | |
| 1. štjax da săm pisal | štjal săm da săm pisal | štjal săm da săm pisal | štjal săm **bil** da săm pisal |
| 2. šteše da si pisal | štjal si da si pisal | štjal si da si pisal | štjal si **bil** da si pisal |
| 3. šteše da e pisal | štjal **e** da e pisal | štjal da e pisal | štjal **bil** da e pisal |
| 1. štjaxme da sme pisali | šteli sme da sme pisali | šteli sme da sme pisali | šteli sme **bili** da sme pisali |
| 2. štjaxte da ste pisali | šteli ste da ste pisali | šteli ste da ste pisali | šteli ste **bili** da ste pisali |
| 3. štjaxa da sa pisali | šteli **sa** da sa pisali | šteli da sa pisali | šteli **bili** da sa pisali |
| *Aorist* | | | |
| 1. pisax | pisal săm | pisal săm | pisal săm **bil** |
| 2. pisa | pisal si | pisal si | pisal si **bil** |
| 3. pisa | pisal **e** | pisal | pisal **bil** |
| 1. pisaxme | pisali sme | pisali sme | pisali sme **bili** |
| 2. pisaxte | pisali ste | pisali ste | pisali ste **bili** |
| 3. pisaxa | pisali **sa** | pisali | pisali **bili** |

Figure 5.2: Forms of the verb *piša* ‘to write’, part 2

### 5.2.2 History of description

Interestingly, hearsay and inferential forms have different status in Bulgarian traditional grammar[9]. Hearsay forms have been noted since the second half of the XIXth century. The first comprehensive grammar that considered the whole paradigm of these forms is Andrejčin (1944). In that grammar, they are simply called *Renarrated tenses*. In more recent grammars, such as Bojadzhiev et al. (1983) and Scatton (1984), they are listed under the name *Renarrated mood*. Each of these grammars devotes a separate section to hearsay forms.

Inferential forms, on the other hand, are scarcely mentioned in Scatton (1984). Bojadzhiev et al. (1983) has a half-page footnote (p. 324) in a section on Perfect, where the inferential uses of Perfect are explained and the existence of a whole subsystem of forms parallel to Renarrated is briefly mentioned. Andrejčin (1944) notes it in separate paragraphs within his discussion of the renarrated forms that retaining the auxiliary signals 'that actions not immediately witnessed by us are related as our personal statement' – in his discussion of Aorist (§292), Imperfect (§295), Pluperfect (§300), and Future Perfect (§306).

### 5.2.3 Friedman's objections to the mainstream analysis

While the analysis of the Bulgarian verb that includes both Renarrated and Conclusive as separate forms has now become mainstream among linguists in Bulgaria (and Russia)[10], there remains considerable opposition to

[9] In discussing the history of treatment for evidential forms in the Bulgarian tradition, I am following Kutzarov (1984, 1994) and Nitzolova (2007).

[10] Important differences still exist among Bulgarian linguists as to the number of evidential forms (such as whether Conclusive can be used in present and future tenses), the

it in the Western tradition. Perhaps the most vocal among those opponents is Victor Friedman, whose numerous works maintain that perfect-based forms in Bulgarian, as well as in other languages (Macedonian, Albanian, Turkish, Georgian and more) should not be taken as expressing evidentiality.

His position can be summarized as follows:

1. Forms traditionally treated as evidential in Bulgarian and Macedonian do not in fact express evidentiality as part of their grammatical meaning ("These forms are not special evidential forms but rather forms capable of expressing evidentiality", Friedman (1986, p. 169)).

2. There is no grammatical distinction between the 3rd person forms with retained auxiliary and elided auxiliary in spoken dialects. This is entirely determined by pragmatics.

3. The tradition distinguishing Renarrated and Conclusive is a result of L. Andrejčin's immense influence. "...one is tempted to suggest that, like Noam Chomsky in the United States or Nikolaj Marr in the Soviet Union, Ljubomir Andrejčin was a linguist engaged in a power struggle for the hegemony of his ideas and analyses, and like those other linguists, Andrejčin was successful.[11]" (Friedman, 2002, p. 209)

role of the Dubitative (whether it should be treated as the Conclusive of the Renarrated (Kutzarov, 1994)), whether both or either of Renarrated and Conclusive should be treated as grammemes in the Mood category, and other questions.

[11] Friedman makes an (irrelevant for us now) factual mistake when he says "In Marr's case, the victory only lasted as long as Stalin". Marr's pseudo-scientific theory, after years of hegemony in the Soviet Union in the 30-s and 40-s, was in fact denounced by Stalin himself, in his 1950 article 'Marxism and the problems of language study' (*Marksizm i voprosy jazykoznanija*).

4. The fact that in standard language forms with and without the auxiliary are used in different ways is due to prescriptive pressure. Here Friedman makes a comparison to the definite article morphemes in standard Bulgarian, which combine forms from different dialects and introduce case distinctions which are absent from every particular dialect.

There is a lot of interesting evidence in Friedman's works, and it deserves careful consideration. Here are some counterarguments:

1. Even if the grammar of standardized Bulgarian differs from grammars of individual dialects, it is still a grammar of a natural language. Similar situations exist almost everywhere standard languages exist; sometimes the difference between spoken dialects and a standard language is considerably larger than in the case of Bulgarian (for example, in Arabic or Chinese). Still, speakers of the standard language have intuitions and these intuitions are worth investigating. Speakers I worked with demonstrate considerable agreement as to where Renarrated and Conclusive are appropriate, even though no standard grammar text specifies where the distinction lies with enough precision.

   Friedman himself notes that many of his problematic examples, even though they are taken from naturally occurring spoken discourse, are judged ungrammatical by the majority of the speakers.

2. It is not true that Andrejčin's analysis is considered infallible by Bulgarian grammarians. Andrejčin (1944) does not have a detailed description of Conclusive (even though, according to Kutzarov (1994, p. 30), his 1936 doctoral dissertation is closer to modern views of scholars like Kutzarov

himself, Gerdzhikov or Nitzolova). It is also not true that earlier descriptions by Tzonev (1910) contradict Andrejčin's judgments, as claimed in Friedman (2002, p. 216). Considering example

(69) (a) Rekata pridošla.

(b) Rekata e pridošla.

'The river has risen'

Tzonev identifies (2a) as Renarrated (without using the term) and assigns it greater confidence than the Conclusive (2b). Andrejčin (1944, 292) calls Conclusive 'expression of a past action as a personal statement, even if we have not witnessed it'. Indeed, Conclusive signals 'personal' inference on the part of the speaker, but there is no guarantee that this personal statement expresses greater confidence than that of a reliable witness. See discussion of (Fitneva, 2001) below.

3. Some of Friedman's examples are problematic; some others are explainable within the boundaries of the mainstream theory.

(70)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Obadix se | na | čičo | mi. | Ne | beše | văv | kăšti, |
| call.Aor.**I**.1Sg | to | uncle | my | not | be.Aor.**I**.3Sg | in | house |

| | | |
|---|---|---|
| bil e | na | plaža |
| be.Aor.**C**.3Sg.M | on | beach.Def |

'I called my uncle. He wasn't home, apparently he was at the beach'[12]

[12] The original example from (Friedman, 2000, p. 331) is in Macedonian (*Mu se javiv na vujko mi. Ne beše doma, na plaža bil.*). The Bulgarian translation exhibits the same distribution of evidential markers.

Here, strictly speaking, the speaker obtained all information over the phone from his aunt. But the information that the uncle was not home was received from a more immediate (good enough) source, so there are sufficient grounds to use the Indicative. On the other hand, information that the uncle was at the beach was only obtained through hearsay; the speaker's aunt could not see whether it was true. So using Renarrated or Conclusive is required (my consultant preferred the Conclusive, given that the aunt is a reliable source). Of course, directness of information source correlates with its reliability, which is the parameter that Friedman uses to explain this example[13].

4. Certain features of Bulgarian forms as described by the mainstream linguists find parallels in descriptions of unrelated languages with evidentials; thus they may represent instances of typologically significant generalizations (whether universals or tendencies). One would not expect such regularities if the evidential nature of forms involving *l*-participle were just a figment of prescriptive linguists' imagination.

    - Renarrated forms have wider applicability than indirect speech.
    - There are narrative genres (specifically, folklore tales) where Renarrated is used almost exclusively.
    - Significant events that occurred some time ago may be narrated in Indicative even when the speaker was not a witness to them. See discussion in Section 5.2.5.

[13] Martina Faller (p. c.) objects to my attempt to use the notion of the best possible source (Faller, 2002) here, since in principle the event of the uncle's absence could be witnessed personally by someone in the house (e. g., the speaker's aunt).

This example also poses a problem for Koev's theory of the meaning of Indicative, discussed in §5.3.3, see p. 146.

5. As an explanation for the rules of omitting the third person auxiliary in the forms with an *l*-participle, Friedman refers to the works of Grace Fielder, such as (Fielder, 1995). Fielder maintains that the main parameter that influences the choice of the form is whether the information in the sentence is foregrounded or backgrounded. In some of her examples this indeed seems to be the best explanation. However Fielder herself admits that not every case of auxiliary omission or retainment can be explained in this way. Moreover, she recognizes that there is high correlation of AUX+ — her code for *l*-participle forms with the auxiliary, — with, first, inference on the part of the speaker himself as opposed to mere hearsay (cf. Demina (1959); Kutzarov (1994); Nitzolova (2007) and many others) — many of her examples involve explicit markers of epistemic modality or supposition, and, second, with personal statements of the speaker (cf. Andrejčin (1944, §§292,295,300,306)). In my experience, speakers tend to prefer the Conclusive when they consider the source of hearsay information to be reliable and are ready to take responsibility for its veracity.

Friedman (1986) interprets forms traditionally labeled as Indicative mood as expressing confirmation on the part of the speaker; forms with pluperfect marking (Dubitative in our terminology) as expressing lack of confirmation, and forms with perfect marking (both Renarrated and Conclusive, according to our terminology) as neutral with respect to speaker confirmation. He further claims that there is no systematic difference in meaning between forms where the auxiliary is omitted in the 3rd person ('Renarrated') and those where it is retained ('Conclusive').

A compelling case against Friedman's (and others) analysis is provided by Fitneva (2001). The main idea of that article is to test whether Bulgarian grammatically encodes the source of information (as the mainstream Bulgarian grammarians maintain) or the speaker's degree of certainty (Friedman's theory clearly falls into this class). Fitneva notes that for different classes of situations either hearsay evidence or inference from indirect evidence can provide greater certainty.

> . . . In the context of a story, a protagonist A asked the characters B and C about their friend D. B and C gave different information and, crucially, their statements were worded using different word endings. The children's task was to say whom they think A believed.. . .
>
> To distinguish between the evidential and speaker attitude theories, in the stories used in the first study, A asked about D's whereabouts; in the stories used in the second study, A asked what D did. The prediction was that asking about location would strongly bias children to seek perceptually acquired information; and how the information was acquired will be more important than whether it was speaker or someone else who acquired it, i. e., the [−BE] perfect would be judged more reliable than the [+BE] perfect. This need not be the case when people ask about actions. Friedman's theory, on the other hand, predicts that in both cases the [+BE] and [−BE] perfect will be judged equally reliable, because the two forms are equivalent in expressing the non-confirmation of the speaker.
>
> (Fitneva, 2001, p. 411)

Here is an example text (ibid., p. 412):

The turtle and the hedgehog are the rabbit's best friends. One day they decided to go to a movie. On the way, they met the snail.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ako | njakoj | znae, | neka | da | mi | kaže | kăde | da | namerja |
| if | someone | knows | let | to | me | tell-3p-SG | where | to | find-1p-SG |

| |
|---|
| Zajo |
| rabbit |

'If you know please tell me where to find the Rabbit,' he said

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Zajo | e | otišăl | da | si | počine | pod | starija | dăb. |
| Rabbit | is | gone | to | REFL | rest | under | old | oak |

'The Rabbit must have gone to take a nap under the old oak,' said the turtle.

| | | | | | | |
|---|---|---|---|---|---|---|
| Zajo | otišăl | da | risuva | pri | golemite | skali. |
| Rabbit | gone | to | draw | by | big-THE | rocks |

'The Rabbit, someone said, went to paint by the big rocks,' said the hedgehog.

| | | | | | |
|---|---|---|---|---|---|
| Na | kogo | li | e | povjarval | ohljuvăt? |
| to | whom | PARTICLE | is | believed | snail-THE |

'To whom did the snail believe?'

9-year old children in Fitneva's experiments showed the predicted behaviour (although to a slightly less than statistically significant degree): when asked about a character's whereabouts, 75% indicated [−AUX] form as more reliable; when asked about actions, 67% preferred the [+AUX] form. In 6-year old children, the tendency was present as well, although much weaker.

It is important to note that Fitneva's examples do not demonstrate any foregrounding-backgrounding contrast, so that Fielder's theory is unable to explain the observed distinctions. Neither is it likely that 9-year old children are under strong pressure of prescriptive norm, as Friedman would have it. One is therefore justified to conclude:

- speakers of Bulgarian distinguish between constructions with and without the auxiliary (thus, between Renarrated and Conclusive forms);
- depending on the type of situation, either information presented in Conclusive or Renarrated is judged as more reliable. This confirms the hypothesis that these constructions mark information source, and not just degree of confirmation.

### 5.2.4 Homonymy

The same morphological form often occurs in several cells of a verb paradigm; in many cases, different evidential forms are not distinguished, in others, evidential forms coincide with certain non-evidential ones. In each case, we need to decide whether we deal with homonymy, or whether there is an invariant meaning that can be attributed to certain forms.

First, as I mentioned already, within each type of evidentials tenses with the reference point in the past are not distinguished from those with the reference point in the present. Here, an invariant (temporal and aspectual relation with respect to the reference point) is easy to provide; on the other hand, since the existence of Conclusive forms for the present and the future is in doubt, and Renarrated forms have different markedness status in the past and present tenses as well (see Section 5.2.5), it may be advantageous to consider these tenses separately.

Second, in the first and second person, Renarrated forms are not distinguished from Conclusive. One can consider these forms instances of a more general Indirect evidential. The first and second person are highly marked forms for indirect evidentials: it is uncommon to have indirect evidence about oneself, or to notify one's interlocutor about some indirect information concerning him/her. Most uses of indirect evidentials are in the 3rd person. It is thus not surprising that certain distinctions present in the unmarked forms are neutralized in the marked forms.

Third, Aorist for Conclusive has the same form as Indicative for Perfect. In this case, it is hard to find any invariant meaning; some instances can clearly only be analyzed as Perfect Indicative, some as Aorist Conclusive. However many instances can be treated both ways, so a grammarian has to adopt some arbitrary decision on how to classify them (for example, Nitzolova (2007, p. 135) considers all the problematic examples Conclusive unless the event in question was clearly witnessed by the speaker). One class of such problematic cases is indirect speech, where Perfect is the preferred tense used to refer to past events.

Fourth, Renarrated forms in Aorist and Present coincide with Mirative — a special set of forms used to express information surprising for the speaker. While having the same construction express indirect evidentiality and mirative is common enough among languages (Turkish and Albanian are cases in point; in both languages, the construction is historically derived from the perfect as well), it is probably impossible to find a meaning component that unites the two. Slobin and Aksu-Koc (1986) speak about "unprepared mind of the speaker" as an invariant in Turkish. However, in Bulgarian, some uses of Renarrated do not express unprepared mind. Moreover, many uses of Conclusive, where an inference has just been made, and the proposition expressed is not yet added to the speaker's core belief set, would be better examples of unprepared mind than hearsay cases.

Fifth, the difference between Imperfect/Present evidential forms and Aorist forms is that a different kind of *l*-participle is used (for *piša* 'write', the Imperfect participle is *pišel*, the Aorist participle *pisal*). This participle is a relatively recent innovation, and not all dialects of Bulgarian have it. Even in those dialects that do, the participles coincide for some verbs, including those of a productive IIIrd conjugation (for example, the verb *strelja* 'to shoot' has *streljal* as both Aorist and Imperfect *l*-participle). Thus, many verbs do not distinguish Aorist and Present/Imperfect evidential forms (Nitzolova, 2007, p. 145).

Such high degree of homonymy may appear strange, but it is far from being exceptional. One can compare this to a list of interpretations for almost any case in almost any nominal case system (take Ablative in Latin, Instrumental in Russian or Genitive in any language). Another example is the multitude of uses of the *-ing* suffix in the English morphology: it serves to denote at least the main verb in Progressive tenses, participles, gerunds, and deverbal nouns.

### 5.2.5 Indicative

As noted by Stankov (1969, p. 166), Bulgarian past tenses in Indicative (Aorist and Imperfect) have acquired presupposition of speaker's direct evidence as part of their grammatical meaning. As a result, when speaking about events where direct evidence is absent, the speaker is *forced* to use one of the forms of indirect evidentiality. On the other hand, Bulgarian Indicative Present and Future are neutral with respect to the source of information. Thus, when the source of the information is hearsay, the speaker has a choice between Renarrated and Indicative. Present Renarrated then acquires a pragmatic connotation of disbelief.

What evidence counts as direct depends on the genre of the text and the proposition being expressed. For everyday events, the speaker needs to be a direct witness. For medium-scale historical events, the event has to happen within the speaker's lifetime, and the speaker needs to be aware of the news when they happen. My consultant, who was born in the 60-s, cannot use Indicative to say

(71) Stalin umrja prez 1953 g.
S. die.Aor.**I**.3Sg in 1953 y.
'Stalin died in 1953'

but his parents can, even though at the time they got the news from hearsay. Similar change of perspective with time has been described by Slobin and Aksu-Koc (1986, p. 163) for Turkish. See also the discussion of this example in §5.3.3.

In history books, major distant historic events can be introduced by Indicative, if they are well known, but detailed description will proceed in Renarrated or Conclusive.

In the language of the press, the journalist can use Indicative or Renarrated to demonstrate, respectively, smaller or greater distance from the proposition he relates. During the Communist rule, information from the Soviet news agency TASS was reported in Bulgarian newspapers using Indicative, while information form Western news agencies was reported in Renarrated (Stoichkova and Chausheva, 1995, p. 259).

### 5.2.6 Renarrated

Renarrated is used to indicate that the information reported by the speaker is based on hearsay. The speaker does not guarantee the veracity of this information. In the past tense, Renarrated does not carry any indication whether the speaker trusts the information; however, in my tests, when they do trust it, in the present and future tenses, Indicative tends to be used, and in the past tenses, Conclusive. The use of Renarrated in everyday speech thus pragmatically implicates some degree of disbelief. This implication is absent in genres such as history books, where Renarrated is required for distant past events. Fairy tales are also told using Renarrated.

In the rest of this section I will be concerned with the question: what exactly is considered hearsay information?

#### 5.2.6.1 Multiple sources

The information contained in a Renarrated clause may come from several unrelated past utterances. Consider the following example:

(72) *Ivan: Konstantin used to live in Varna, but two years ago he moved somewhere else.*
*Peter, on an unrelated occasion: Konstantin used to live somewhere else, but two years ago he moved to Sofia.*

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| Boris: | Predi | dve | godini | Konstantin | **se premestil** | ot | Varna |
|  | ago | two | years | K. | move.Aor.**R**.3Sg.M | from | V. |

|  |  |
|---|---|
| v | Sofia. |
| to | S. |

‘Two years ago Konstantin (reportedly) moved from Varna to Sofia.’

In Boris’s utterance, Renarrated is used, even though two primary sources are combined, and neither of these sources possesses all the information provided in the Renarrated utterance.

The fact that multiple sources are allowed in building a Renarrated utterance reduces the cognitive load on the speaker — one does not have to trace the source of every piece of second-hand information that one possesses; it is sufficient to mark the information as second-hand. In turn, this allows the use of Renarrated in folk tales and history texts, where multiple sources are combined on a regular basis.

On the other hand, since everyone can easily remember a pair of past utterances that contradict each other, the notion of information as used in Renarrated cannot be possibly closed under entailment[14] — otherwise one could have used this contradiction to justify an arbitrary statement in Renarrated.

#### 5.2.6.2 Degree of inference

Although Renarrated can be compared to indirect speech, the speaker who uses Renarrated has more freedom in choosing the representation of the information. For example, if A said:

(73) John has sold a car to Bill.

[14] At least classical entailment where contradiction implies anything.

this arguably does not constitute grounds for the statement:

(74) A said that Bill has a car now.

On the other hand, if A says

(75) Vasil sold a car to Angel yesterday

this does constitute grounds for

(76) Angel imal kola
A. have.M.Sg.**R**.Pres car
'Angel (reportedly) has a car.'

Scalar inference is possible:

(77) *Boris: I solved three problems at the exam.*

Boris rešil po-malko ot pet zadači
B. solve.Aor.**R**.3Sg.M less from five problems

'Boris (reportedly) solved less than five problems.'

The following several examples test how far the reported proposition can diverge from the text of the report.

Suppose Peter tells Ivan:

(78) Kartofite sa poskăpnali dva păti
potatoes.Def rise.in.price.Perf.Pl two times
'Potatoes have become two times more expensive.'

If Ivan knows that the old price was 0.5 lv., he can then say

(79) [Petr kaza če] kartofite stanali edin lev
Peter said that potatoes.Def become.Aor.**R** one lev
'[Peter said that] potatoes have (reportedly) become 1 lv.'

(This example is based on (Faller, 2002, p. 271).) Ivan does not have to keep track of exactly what information comes from Peter; there is even some room for misunderstanding: perhaps Ivan thought that the old price was 40 st., so the new one is actually 80 st.

Another example (also based on (Faller, 2002, p. 271)). Penka tells Minka the day before the party:

(80) Utre šte xodja v Sofia
tomorrow go.Fut.1Sg to S.
'Tomorrow I will go to Sofia.'

in this case, Minka *cannot* use Renarrated in the following way when discussing the party which is supposed to take place at a time when Penka is in Sofia:

(81) # Penka njamalo da dojde
P. come.Fut.Neg.**R**
'Penka (reportedly) will not come.'

(Faller reports the hearsay evidential *is* possible in Cuzco Quechua in such a setup.)

A third example (this time based on (Matthewson et al., 2007, p. 16)). Suppose Ivan looks at a bottle and says:

(82) Viždam dve šišeta
See.Pres.1Sg. two bottles.Def
'I see two bottles.'

Even though this can be considered evidence that Ivan is drunk, one cannot say

(83) # Ivan bil pijan
I. be.Pres.3Sg.**R** drunk
'Ivan (reportedly) is drunk.'

(Matthewson et al. report that the hearsay evidential is also impossible in St'át'imcets.)

The distinguishing feature between the cases where Renarrated is allowed and those where it is not seems to be the following: in the examples where Renarrated is possible, the original speaker has the intention to induce belief in the proposition that is reported. In (75), C wants his interlocutor to believe that the car now belongs to Angel. In (78), Peter wants to make Ivan believe that the new price is 1 lv (or at least, in the case of confusion, Ivan takes Peter to have that intention). On the other hand, in (80), Penka does *not* have the intention to make Minka believe that she will miss the party, even though she does herself have this belief and acquisition of this belief by Minka is one of the results of her utterance. In (82) Ivan has no intention to make anyone believe he is drunk; he just reports what he sees. On the other hand, if Ivan does in fact know that there is only one bottle, and uses his utterance as a way to indicate that he is drunk, (83) becomes appropriate.

The report need not be based on a single utterance from a single source. On the other hand, it cannot rely on the absence of certain utterances in a large corpus. In other words, derivation operations allowed in producing a Renarrated statement do not include inference from ignorance. The statement in (84) follows from a corpus of hearsay knowledge of Bulgarian history, combined with the assumption that this corpus covers all the important relevant events. Still, Renarrated is not justified in this example.

(84) Ungarcite nikoga ne *(sa) zavladjavali Bălgaria
Hungarians never not conquer.Aor.**C**/#**R**.3Pl Bulgaria
'Hungarians never (apparently/#reportedly) conquered Bulgaria'

### 5.2.7 Conclusive

In many cases, Conclusive is used in situations where the proposition stated in the sentence is known by inference. The inference in question can be of any type:

- From results to the process that led to those results.

  (85) *Situation: there are empty bottles in Ivan's room.*

  Ivan e pil vino včera
  I. drink.Aor.**C**.3Sg.M wine yesterday

  'Ivan drank wine yesterday.'

- From a process to the result of this process.

  (86) *Situation: Ivan went home at 8pm.*

  Ivan se e pribral predi 10 č.
  I. Refl return.Aor.**C**.3Sg.M before 10

  'Ivan returned (home) by 10pm.'

- From a general statement to a particular instance.[15]

[15] Izvorski (1997) states that this is not possible. Her example:

(87) *Situation: Ivan always goes to work on Tuesdays.*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Ivan | e bil | na | rabota | văv | vtornik | văv | 2 č. |
| I. | be.Aor.**C**.3Sg.M | at | work | in | Tuesday | in | 2 |

'Ivan was at work on Tuesday at 2pm.'

(88) *Knowing how much Ivan likes wine...*

| | | | | | |
|---|---|---|---|---|---|
| toj | sigurno | e izpil | vsičkoto | vino | včera |
| he | surely | drink.Aor.**C**.3Sg.M | all.Def | wine | yesterday |

'Ivan surely must have drunk all the wine yesterday.'

Here *sigurno* 'surely' indicates that the defeasible inference lacks reliability.

- Inductive inference.

(89) *Situation: three eggs in a carton turned out to be rotten. Speaker does not want to check the rest of them.*

| | | |
|---|---|---|
| Vsičkite | jajca | sa se razvalili |
| all.Det | eggs | rotted.Aor.**C**.3Sg.Pl |

'All the eggs have rotted.'

- Inference to the best explanation.

(1) *Knowing how much Ivan likes wine...*

| | | | | |
|---|---|---|---|---|
| # toj | izpil | vsičko | vino | včera |
| he | drink.Aor.**C**.3Sg | all | wine | yesterday |

'Ivan (apparently) drank all the wine yesterday.'

Note, however, that Izvorski does not use the auxiliary in this example. So, according to the descriptions I have, this would count as Renarrated.

(90) *Situation: Ivan should have arrived some time ago, but there is no sign of him. It is known that there have been disruptions in the bus schedule.*

Ivan sigurno e vzjal avtobusa
I. surely take.Aor.**C**.3Sg.M bus.Def

'Surely Ivan took the bus.'

- Mathematics.

(91) Petăr otival ot A do B za 2 časa
P. go.Impf.**R**.3Sg.M from A to B in 2 hours
'Peter used to go from A to B in 2 hours.'

Ot B vednaga trăgval za C, dokădeto
from B immediately go.Impf.**R**.3Sg.M to C where
pătuval ošte 1 čas
travel.Impf.**R**.3Sg.M more 1 hour

'From there he would immediately start off to C, and traveled there in 1 more hour.'

Znači ot A do C e stigal za 3 časa
means from A to C reach.Impf.**C**.3Sg.M in 3 hours

'Therefore, he used to reach C from A in 3 hours.'

In this example, I use Imperfect in order to clearly distinguish Conclusive from Indicative Perfect.

(92) *In a history book, following a list of possessions in a document from 1762.*

Edna puška e struvala kolkoto edna krava s tele
one gun cost.Aor.**C**.3Sg.F as much as one cow with calf
— tvărde visoka cena, tăj kato oružieto po onova vreme
very high price as weapons.Def in that time
se e cenjalo
be.valued.Impf.**C**.3Sg.N

'A gun (apparently) cost as much as a cow with a calf: a very high price, since weapons were (apparently) valued very much at that time.' (Taxov, cited by Nitzolova (2007, p. 177))

The last type of inference is especially interesting. The passage comes from a text where Renarrated would be expected for information taken from the documents. Nitzolova remarks that the passage presents the historian's conclusions based on the document. Since in this case premises strictly imply the conclusion (at least the clause with the verb *cost*), there is no room to construct a possible worlds structure where everything in the document were true and the statement about costs were false. Therefore, from the fact that Renarrated was not chosen for this verb, we can conclude that Renarrated in Bulgarian cannot be analyzed as an epistemic modal, the way Izvorski (1997) (for Bulgarian) and Matthewson et al. (2007) (for St'át'imcets) do.

However there are uses of Conclusive where no inference is involved. Nitzolova (2007) states that Conclusive can be used in cases of the so called "weak knowledge" of the speaker — knowledge that is not based on speaker's own perception. Thus it can be used both for inference and for hearsay knowledge:

(93) Ošte v detstvoto si majka mi e svirela na tsigulka
already in childhood her mother my play.Impf.**C**.3Sg.F on violin

'My mother played violin when she was a child.' (Nitzolova, 2007, p. 138)

The difference between Renarrated and Conclusive in such cases is that in cases of Renarrated it is part of the grammatical meaning that the information is based on hearsay, and speaker's attitude to the information is not expressed. On the other hand, if Conclusive is used, the grammatical meaning tells us that the speaker takes responsibility for the veracity of expressed information, and the source is not expressed (other than that it's not perception). The general meaning of Conclusive seems to be that of indirect evidentiality, rather than that of an inferential marker.[16] Nitzolova provides examples where both Renarrated and Conclusive are used alongside each other:

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| (94) | Djado | mi | njakoga | kato | učitel | e igral | Ivanko, | a |
|  | grandfather | my | once | as | teacher | play.Aor.**C**.3Sg.M | Ivanko | and |
|  | po-kăsno | stanal | sveštenik |  |  |  |  |  |
|  | later | become.Aor.**R**.3Sg.M | priest |  |  |  |  |  |

'My grandfather, being a teacher, (must have) played (the part of) Ivanko, and (reportedly) later became a priest.'

### 5.2.8 Evidentials in embedded contexts

Evidential markers cross-linguistically have the tendency to only occur in the main clause of a sentence. This is not the case with Bulgarian.

Both Renarrated and Conclusive are used in indirect speech or as complements of verbs of propositional attitude. In this case, Renarrated meaning seems to be redundant, since it is already conveyed by the speech/propositional

[16] Stankov (1969, p. 174) considers these forms neutral with respect to source of information, as opposed to past Indicative (direct evidence required) and Renarrated (hearsay).

attitude verb. The choice between Renarrated, on one hand, and Conclusive or Indicative (depending on tense), on the other hand, implicates degree of speaker's belief in the hearsay information conveyed (see also §5.3.2).

Renarrated and Conclusive can also occur in other types of subordinated clauses, such as antecedents of conditionals.

(95)

| | | | | | |
|---|---|---|---|---|---|
| Ako | izlezel, | šteli$_1$ | da$_2$ | go | vidjat$_3$ |
| if | come.out.Aor.**R**.3Sg.M | | | him | see.PFut.**R**.3Pl$_{1,2,3}$ |

'If he came out, they would (reportedly) have seen him' (The speaker describes the situation on the basis of someone else's words.)

Not all of my consultants accept (95) as grammatical.

(96)

| | | | | | | |
|---|---|---|---|---|---|---|
| Ako | sa moželi | da | pišat, | trjabva | da | ima |
| if | can.Impf.**C**.3Pl | to | write.Pres.3Pl | must | to | have.Pres.3Sg |

| | |
|---|---|
| tekstove | njakăde |
| texts | somewhere |

'If they could (apparently) write, there must be texts somewhere.'

Note that in (95), both the antecedent and the consequent of the conditional are in Renarrated; one can probably say that the antecedent agrees with the consequent in its evidential marking. In (96), on the other hand, only the antecedent has Conclusive marking. This example clearly shows that the use of the evidential is descriptive, not m-performative, according to the terminology of Faller (2006), and thus does contribute to the propositional content of the sentence.

Evidential markers can be used within a restrictive relative clause:

(97)

| | | | | |
|---|---|---|---|---|
| Vidjax | čovek | koito | ograbil | banka |
| See.Aor.1Sg | man | which | rob.Aor.**R**.3Sg.M | bank |

'I saw a man who (reportedly) robbed a bank.'

The indefinite NP containing this relative clause can have narrow scope with respect to other operators in the sentence:

(98) Ako vidja čovek koito ograbil bankata, šte
if see.Pres.I.1Sg man which rob.Aor.**R**.3Sg.M bank.Def Fut
se obada na policiyata
call.1Sg on police.Def
'If I see a man (anyone) who (reportedly) robbed the bank, I will call the police.'

One distinctive feature typical of evidentials is that they always take wide scope with respect to negation. This holds for Bulgarian as well:

(99) Ivan ne zaminal
I. Neg leave.Aor.**R**.3Sg.M
'Ivan (reportedly) didn't leave.'

does not have the reading 'I don't have hearsay information that Ivan left'.

## 5.3 Formal theories of evidential meaning

There has been a lot of work in formal semantics in recent years on the topic of evidentials. While most of the problems that attracted the linguists' attention are orthogonal to my concerns, a short survey is in order.

### 5.3.1 Evidentials as not-at-issue meaning

Perhaps the most widely discussed property of evidentials is lack of interaction with the main content of the sentence containing them. In particular, evidentials are never directly challengeable[17]:

[17] I illustrate with Bulgarian examples.

(100)

| — Ivan | izpil | mnogo | vino |
|---|---|---|---|
| I. | drink.Aor.**R**.3Sg.M | much | wine |

‘Ivan reportedly drank a lot of wine.’

| — Ne | e | vjarno! |
|---|---|---|
| Not | is | true |

‘That’s not true!’ (can only mean ‘He did not drink a lot of wine!’, not $^{\#}$‘You did not hear that!’)

In most languages, evidentials appear to take wide scope with respect to operators such as negation.

(101)

| — Ivan | ne | izpil | mnogo | vino |
|---|---|---|---|---|
| I. | not | drink.Aor.**R**.3Sg.M | much | wine |

‘Ivan reportedly didn’t drink a lot of wine.’, *not* ‘$^{\#}$ I didn’t hear that Ivan drank a lot of wine’

As we can see, the evidential meaning cannot be part of the sentence assertion. Several theories have been proposed as to what it is:

- evidential meaning as presupposition (Izvorski, 1997; Matthewson et al., 2007; McCready and Ogata, 2007);
- evidentials as illocutionary operators (Faller, 2002);
- evidentials as not-at-issue assertion (Murray, 2010; Koev, 2010).

#### 5.3.1.1 Evidential meaning as presupposition

Izvorski (1997) proposed a theory that treats evidentials as a kind of epistemic modal, using Bulgarian for most of her examples. She couches her analysis in terms of Kratzer’s theory of modality (Kratzer, 1981). The modal base is taken to be the set of worlds where indirect evidence is true (premises of inference for inferential evidentials, existence of the report for hearsay evidentials). The ordering source is normality — worlds are more normal where the evidence actually points to the expected conclusions and reports are true.

Having established what kind of modality is expressed by evidentials, Izvorski presents the following analysis for $p$-**R**:[18]

- Assertion: $\Box p$ *in view of speaker's knowledge state*

- Presupposition: *speaker has indirect evidence for p*

One problem with this analysis is that presuppositions typically can be satisfied at levels other than the outermost discourse scope. For example, it should be possible to bind the hearsay presupposition in the conditional antecedent.

For Conclusive, presupposition cancellation does indeed work:

(102)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Ako | v | kabineta | na | Ivan | ima | šišeta, | znači | včera |
| if | in | office.Def | Prop | I. | have.Pres.3Sg | bottles | means | yesterday |

| | |
|---|---|
| e pil | vino |
| drink.Aor.**C**.3Sg.M | wine |

'If there are bottles in Ivan's office, then he (apparently) drank wine yesterday.'

But for Renarrated, attempts to cancel the presupposition fail.

[18] In that article, both hearsay and inferential evidentiality is taken to be expressed by an *l*-participle without auxiliary, as opposed to Perfect, expressed by an *l*-participle with auxiliary. Thus, on the descriptive level, Izvorski does not agree with most Bulgarian linguists.

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| (103) | Ako | Maria | kazva | če | Angel | kupil | kolata, | $^{\#}$togava |
|  | If | M. | say.Pres.**I**.3Sg | that | A. | buy.Aor.**R**.3Sg | car.Def | then |

|  |  |  |
|---|---|---|
| Angel | imal | kola |
| A. | have.Pres. |  |

‘If Maria says that Angel bought a car, then Angel (reportedly) has a car.’

McCready and Asher (2006) use the mechanisms of Structured Discourse Representation Theory (SDRT) to express the restriction as to where the presupposition is allowed to be bound. Presupposition generated by evidentials, according to their approach, has to be *externally anchored*, which guarantees that it is only satisfied at the highest level of discourse.

Still, one feature distinguishes evidential meaning from most presupposition types. Presupposition is typically expected to be satisfied in the common ground of the discourse prior to the utterance of a sentence that carries the presupposition. Mechanisms for presupposition accommodation do exist, but their use is exception rather than the rule. Evidential meaning, on the contrary, is typically new information in the discourse context. Thus, if it is presupposition, it is a special kind that is particularly well suited for accommodation.

The analysis of evidentials as epistemic modals fails in cases where there are no possible worlds to play with — such as when the reasoning involved is mathematical. In (91), for example, one cannot use Renarrated for the conclusion, even though it holds in every world compatible with the premises (known through hearsay). One would need to use a non-standard theory of semantics for epistemic modals in order to accommodate this, such as (von Fintel and Gillies, 2009a) (this approach is discussed in more detail in §6.5).

#### 5.3.1.2 Evidentials as illocutionary operators

A theory initially proposed by Faller (2002) treats evidentials as speech act modifiers. An evidential takes a whole speech act as its argument and produces another speech act, specifying its sincerity conditions and illocutionary force.

For example, the hearsay evidential marker *-si* in Cuzco Quechua is analyzed in the following way:[19]

$$-\textbf{si}: \begin{array}{l}\textsc{assert}(p)\\ \textsc{sinc} = \{Bel(s,p)\}\end{array} \longmapsto \begin{array}{l}\textsc{present}(p)\\ \textsc{sinc} = \{\exists s_2[Assert(s_2,p) \wedge s_2 \notin \{h,s\}]\}\end{array}$$

That is, a sentence without evidentiality markers expresses a speech act of assertion, with sincerity conditions stating that the speaker believes the contents of this assertion. When modified by the *-si* marker, it transforms into a speech act where the content of the proposition is just presented, not asserted, and the sincerity conditions state that there is another speaker, other than the current speaker and hearer, who asserted this proposition.[20]

#### 5.3.1.3 Not-at-issue propositions

Still another theory of evidential contribution to meaning is defended in Murray (2010). This approach is based on dynamic epistemic logic, so the meaning of each sentence is taken to be an update function on the common ground of the conversation. In Murray's theory, the effect of an utterance containing an evidential consists of *two* updates: first, a non-negotiable update

[19] This is one of the variants presented by Faller (p. 200). On p. 203 another, more complicated analysis is given. Faller does not make a final choice as to which analysis she prefers.

[20] For such analysis to be complete, one would need to specify how an evidential morpheme transforms a speech act other than assertion, such as questions.

with a proposition that is not at issue and not directly challengeable, and, second, a proposal for a further update that can be rejected by the hearer. The language that serves as a testbed for Murray is Cheyenne.

Here is the analysis for a Cheyenne sentence marked with a (zero) direct evidential morpheme (p. 128–129):

(104) É-hó'tȧhéva-∅ Floyd
3-win-DIR Floyd
'Floyd won, I'm sure'

$$\begin{array}{llll} \lambda p[ & (p = \lambda w[\mathbf{won}(w, \mathbf{floyd})]) \quad \wedge & \mathrm{CRT}(v_{\mathit{0}}, \mathbf{i}, p) \quad \wedge & p(v_{\mathit{0}}) \leq p(v_{\mathit{1}})] \\ \hline & \text{(at-issue proposition)} & \text{(ev.restriction)} & \text{(ill. relation)} \end{array}$$

Here, the first conjunct introduces a discourse referent for the proposition under discussion, the second conjunct restricts the common ground to those possible worlds where the speaker has certain evidence for this proposition, and the third conjunct establishes an ordering over the remaining worlds in the common ground such that the worlds where the proposition is true are ranked higher than those where it is false. It is then up to the hearer to accept the proposition and thereby erase the suboptimal worlds from the common ground.

This theory serves as a background for Koev's work on Bulgarian evidentiality (Koev, 2010).

#### 5.3.1.4 Tests

Faller (2006) suggested a number of tests to distinguish evidentials amenable to presuppositional analysis (at least, those contributing to the propositional content of an utterance) from those that work above the propositional level, as speech act modifiers. She argues that roles played by evidentials differ across languages (and perhaps sometimes even across evidential morphemes within one language). The collection of tests was further extended by (Matthewson et al., 2007). Let us run them against Bulgarian data.

1. **(In)felicity if the embedded proposition is known to be false.** Following Faller (2002), Matthewson et al. assume that the modal base only contains worlds that the speaker considers possible; therefore, no modal can be true whose prejacent is known to be false. In the case of Bulgarian, Atanassov (2010) reports that four native speakers out of four recognized the following sentence as acceptable:

(105)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Marina | kaza | če | Todor | imal | červena | kosa, | no |
| M. | told | that | T. | have.Pres.**R** | red | hair | but |

| | | | |
|---|---|---|---|
| kosata | mu | e | černa |
| hair.Def | his | is | black |

'Marina said that Todor has red hair, but his hair is (in fact) black.'

In (105), the sentence marked by evidential is embedded in an indirect speech clause, so this may not count as a clear enough case. But my consultant also accepted the version without any embedding:

(106)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Todor | imal | červena | kosa, | no | kosata | mu | e | černa |
| T. | have.Pres.**R** | red | hair | but | hair.Def | his | is | black |

'Todor reportedly has red hair, but his hair is (in fact) black.'

although Dubitative is more appropriate in such a situation:

(107)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Todor | bil imal | červena | kosa! | Černa | mu | e | kosata | na |
| T. | have.Aor.**D**.3Sg.M | red | hair | black | him | is | hair | of |

| |
|---|
| Todor. |
| T. |

‘Todor allegedly has red hair! Black is the color of Todor’s hair.’

On the other hand, Koev (2010) reports that (108) sounds contradictory:

(108) Ivan napravil torta, $^{\#}$no torta njama
I. make.Aor.**R**.3Sg.M cake but cake not-existed.**I**
‘Ivan reportedly made a cake, but there was no cake.’

Again, my consultant judges this sentence to be grammatical,

It seems that acceptability of such examples for native speakers depends on how natural the story sounds, so one has to be careful in devising the test passages. In general, since at least some sentences are acceptable, there appears to be no restriction that requires that the prejacent of the sentence marked by Renarrated be considered possible by the speaker.

In order to reconcile these data with Izvorski’s theory, one only has to claim that the modal base for Reportative evidential is allowed to be wider than the current common ground.

2. **(In)felicity if embedded proposition is known to be true.** Matthewson et al. predict that a modal cannot be used if the proposition is known to be true. In Bulgarian, this is certainly not the case. For example, well established facts of history are presented using Renarrated.

3. **(In)direct evidence requirements not cancelable.** Here, actually, both theories converge in their predictions. In neither of them the type of evidence is conversational implicature, so it should not be cancelable. It is not cancelable in Bulgarian, as one would expect.

4. **Indirect evidence requirement not blocked by negation.** Again, both theories predict that the requirement take wide scope over negation,

and in Bulgarian, it does (see (101)). At the same time, McCready and Ogata (2007, p. 169–170) report that certain kinds of evidentials in Japanese can take narrow scope with respect to a certain type of negation.

5. **Challengeability.** Faller (2002) argues that epistemic modals can be directly challenged, and therefore it should be possible to challenge the content of evidentials if they are a kind of epistemic modals. On the other hand, what would such challenge amount to? Suppose we accept Izvorski's analysis of the Renarrated. In this case, the existence of hearsay evidence should not be challengeable, since it does not form a part of the assertion, but a presupposition. Given that the evidence exists, in normal worlds — those where people speak the truth, — it is guaranteed that the prejacent will be true, and therefore, the modal statement will be true as well. There is nothing to challenge.

6. **Interrogative flip.** Evidentials can be used in speech acts other than assertion. In questions, one interpretation that is present in all languages is that the *hearer* is expected to have information about the answer based on the source specified by the evidential.[21] This phenomenon is called the interrogative flip. Garrett (2001) uses the term *origo* to denote the agent whose point of view is reflected in the evidential morphemes.

   The flip seems relatively easy to explain under the illocutionary theory: a question is a request for an answer, which is another speech act. It is this requested act whose specification is modified by the evidential. Under the

[21] In some languages other interpretations are available, see (Faller, 2002, p. 230) for Quechua and (Murray, 2010, p. 113) for Cheyenne.

modal theory it seems harder to explain how one operator (the question) influences the other (the modal) by modifying its modal base. Since evidentials are a kind of indexicals (see the next subsection for that), the question operator appears to be a monster-inducing environment (Anand and Nevins, 2004), where this particular context index is shifted, while all the others stay in place. In any case, the flip is present both in languages where the illocutionary theory provides better analysis of the data and in those where evidentials seem to be a kind of modals.

In Bulgarian, the flip is present in questions. Example from Nitzolova (2007, p. 119):

(109) *Children are asking while listening to a tale.*

| | | | |
|---|---|---|---|
| I | kakvo | stanalo | posle? |
| and | how | happen.Aor.**R**.3Sg.N | later |

'And what happened next?'

Renarrated is also possible in imperatives. In this case, the imperative does not play the role of performative, but is presented as originating from someone else. Again from Nitzolova (2007, p. 118)

(110)

| | | | | | |
|---|---|---|---|---|---|
| ... | Toj | se zaslušva | v | dumite | ⟨...⟩ |
| | he | listen-Pres.I.Refl.3Sg | in | words.Def | |

| | |
|---|---|
| Da okopael | lozeto! |
| spud-DA.**R**.3Sg.M | vine.Def |

'He listens to her words... He is to spud the vine! (Stratiev)

7. **Embeddability.** In the illocutionary modifier theory evidentials should never be embeddable, since they operate on a different level than the

propositional content of the sentence. Similarly, they should not be embeddable on Murray's not-at-issue assertion theory, because there we first apply the not-at-issue content and only then start to work on the at-issue assertion, so there is no way to mix the two operations. In the epistemic modal theory, there is no reason for the content of evidentials not to interact with other propositional content. In Bulgarian we do find such an interaction, as proved by (95)–(97). (97), repeated here for convenience, is particularly interesting, since there is no way to even state what the proposition is in the scope of the evidential: it contains a variable bound in the superordinate clause.

| | | | | | |
|---|---|---|---|---|---|
| (97) | Vidjax | čovek | koito | ograbil | banka |
| | See.Aor.1Sg | man | which | rob.Aor.**R**.3Sg.M | bank |

'I saw a man who (reportedly) robbed a bank.'

### 5.3.2 Speaker orientation

Since evidentials signal the source of evidence that supports the belief in the content of the sentence in their scope, a question arises who the subject of that belief is. In an independent sentence this subject is naturally taken to be the current speaker, but for evidentials embedded in complements of speech or belief verbs, they might have referred to the beliefs of the subject of these verbs.

This is the question investigated by Sauerland and Schenner (2007). They conclude that in Bulgarian indirect speech complements evidentials behave as Kaplanian indexicals, i. e. they are always speaker oriented.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (111) | Maria | kaza | če | Todor | imal | červena | kosa |
| | M. | said | that | T. | have.Pres.**R**.3Sg.M | red | hair |

‘Maria said that Todor has red hair.’

(111) is appropriate when the speaker has not seen Todor’s hair, and it’s his belief that it is red that is based on hearsay; Maria’s belief is based on perceptual evidence. On the other hand, if the speaker has perceptual evidence himself, (112) is appropriate:

(112) Maria kaza če Todor ima červena kosa
M. said that T. have.Pres.**I**.3Sg red hair
‘Maria said that Todor has red hair.’

In contrast, Garrett (2001) states that in Tibetan, embedded reportatives express the source of information from the point of view of the subject of the verb of speech. Hara (2006) reports that interpretation from the point of view other than the speaker of the independent sentence is possible for the Japanese evidential *darou*, not just for speech and propositional attitude verbs, but for clauses introduced by *because* as well.

### 5.3.3 Temporal theory of evidential meaning

In his work, Koev (2010) considers the opposition of Indicative and Renarrated in Bulgarian. His work is best interpreted as clarifying the meaning of Indicative.

Koev’s theory makes use of the following notions: event time, reference time and learning time. Event time is the time fragment when the event described in the prejacent happened. Reference time is the *time frame which the sentence is about*. Learning time is the time when the speaker (or origo, to take care of the cases of the interrogative flip) learned about the event under the appropriate conceptualization.

The main idea of Koev's theory is that Indicative signals that the learning time overlaps with the reference time.

Here are some examples he uses to support his point of view. First, suppose that on Saturday afternoon the speaker learned that Jack went to New York Saturday morning from one of their mutual friends. According to Koev, when asked what Jack did on Saturday morning, the most natural response will be

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| (113) | Džak | otidel | do | Nju | Jork |
|  | J. | go.Aor.**R**.3Sg.M | to | N. | Y. |

'Jack (reportedly) went to New York'

However when asked what Jack did on Saturday (and thus in a context where Saturday afternoon is included in the reference time), Indicative becomes appropriate:

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| (114) | Džak | otide | do | Nju | Jork |
|  | J. | go.Aor.**I**.3Sg | to | N. | Y. |

'Jack went to New York'

On the other hand, suppose the speaker witnessed Nixon erase the Watergate tapes, but only learned what that amounted to some time later (say, from newspapers), outside of the reference time. Then he can describe Nixon's actions using Renarrated:

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| (115) | Kogato | vljazox, | Niksăn | trieše | njakavi | zapisi. |
|  | when | come-in.Aor.**I**.1Sg | N. | erase.Impf.**I**.3Sg | some | records |

'When I came in, Nixon was erasing some tapes.'

|  |  |  |  |  |
|---|---|---|---|---|
| Toj | zaličaval | / | #zaličašete | ulikite. |
| he | remove.Impf.**R**.3Sg.M |  | remove.Impf.**I**.3Sg | the.clues |

‘He was covering up (reportedly) the clues.’

Koev’s theory receives some support from my data. In particular, when the speaker learns about an event at the time it happens, even if only by hearsay, Indicative appears to be appropriate:

(116) *The speaker sits in a room reading. In the next room there is a TV set showing a soccer match. The speaker hears, without looking at the screen: ‘Messi gets the ball... He strikes... Goal!’ Some time later he reports:*

| | | |
|---|---|---|
| Mesi | vkara | gol |
| M. | score.Aor.**I**.3Sg | goal |

‘Messi scored a goal.’

Another phenomenon that is explained by Koev’s theory is the gradual increase in acceptability of Indicative as the event in question recedes further in time. When we talk about events in the distant past, the period that constitutes reference time tends to become longer: we are talking not about the events of March 5, 1953, but events of 1953 as a whole. But of course, the speaker needs to be alive at the time of the event. This explains the peculiar intuitions we see in (71), repeated here for convenience.

(71)

| | | | | |
|---|---|---|---|---|
| Stalin | umrja | prez | 1953 | g. |
| S. | die.Aor.**I**.3Sg | in | 1953 | y. |

‘Stalin died in 1953’ (can only be used by someone who was born before 1953)

(This sentence can be uttered by someone who was alive in 1953, but not necessarily an eyewitness of Stalin’s death.)

At the same time, some details of the distribution between Indicative and other evidentiality markers (Renarrated and Conclusive) remain problematic for Koev.

First, Present Renarrated has to be taken care of. Since in the Present tense the event time $E$ overlaps with the speech time $S$, and the reference time $R$ overlaps with the event time, the only way we can claim that the learning lies in between $R$ and $S$ is to stipulate that $R$ lies entirely in the past — that is, to claim that Present Renarrated is really past, where the event in question spans a longer period than $R$.[22]

Second, one needs to explain examples where different evidential markers are used for events that happen simultaneously, to mark contrast in cognitive distance, as in (70):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (70) | Obadix se | na | čičo | mi. | Ne | beše | văv | kăšti, |
| | call.Aor.**I**.1Sg | to | uncle | my | not | be.Aor.**I**.3Sg | in | house |
| | bil e | na | plaža | | | | | |
| | be.Aor.**C**.3Sg.M | on | beach.Def | | | | | |

'I called my uncle. He wasn't home, apparently he was at the beach'

Koev's theory would predict Indicative marking for both events of the uncle's absence at home and his being at the beach.

Finally, a special provision has to be made for presentation of major events in history books, where Indicative is used.

[22] Friedman (2000, p. 340) makes just such a claim on independent grounds.

## 5.4 Formal presentation

I can see two ways of adapting my basic semantic framework to cover the data presented by the Bulgarian evidentials. (In the discussion that follows, I concentrate primarily on the Renarrated.) One is to assume that the belief box of a speaker who uses an evidential $p$-**R** contains an (internal language) sentence *A said 'q'*, for some primary speaker $A$ and some $q$ such that $p$ is derivable from $q$ using an inference from a given limited class. The other way is to have multiple compartments (similar to the belief box) corresponding to different ways of acquiring information. Thus the speaker of $p$-**R** would simply have $q$ in his compartment used for hearsay information, such that, again, $p$ is derivable from $q$ using an inference from some limited class.

As an example, according to the first theory, when Boris says *Ivan imal kola* 'Ivan (reportedly) had a car', this means that he has something like the sentence *Peter said 'Ivan bought a car yesterday'* in his belief box. According to the second theory, in such a situation Boris has the sentence *Ivan bought a car yesterday* in his hearsay box.

The main advantage of the first theory is that it requires no new machinery. Any credible account of Mentalese will require that it have enough power to represent citations, so any such account will allow a speaker to possess the thoughts that our theory makes him have. The only other requirement is that the agent should be able to perform certain inferences on the complements of these internal citations.

But this theory has trouble explaining markers of *direct* evidence. The use of Renarrated (especially in the past tenses) does not prevent the secondary speaker from believing the content of his utterance. In our example, Boris may sincerely believe *Ivan had a car* and have a representation of this sentence in his belief box. But having this representation will not allow him to say *Ivan imaše kola* 'Ivan had a car', unless he has some perceptual information. So these forms, most frequent in actual speech, would have to represent internal sentences like 'I have direct evidence that $p$' instead of just $p$. In the case of major past events, as in (71), there should be a way to obtain 'I have direct enough evidence that $p$' from 'I have hearsay evidence that $p$' given that $p$ is a far enough event[23].

On the other hand, the theory of a compartmentalized sentence store has something to recommend it. First, we need some compartments anyway, to distinguish beliefs from desires and intentions. So why not add another box for hearsay? A picture arises where sentences are stored in multiple "boxes". Some of those boxes are inside one another (information acquired through the visual channel is a subset of information based on direct evidence; results of logical inference are a subset of beliefs). Every rule of inference should specify which box(es) its premises should come from and which box gets to contain the conclusion.

One problem is that either the number and nature of those compartments will be language specific, or we will need to have as many compartments as there are types of evidentials in the world's languages — which is a lot.

I will now present a formalized analysis of Bulgarian evidential markers using the first theory, i. e., explicit coding of evidential meaning in the internal language.

I use the presupposition theory of evidential contribution. In order to avoid stipulating overwhelming evidence for $p$ as a presupposition in a speech

[23] That is, unless we adopt Koev's theory and claim that Renarrated and Indicative radically differ in the kind of conditions they impose on the speaker's evidence.

act that asserts $p$, I take the presupposition to be 'the speaker has evidence of the right type concerning the proposition $p$' — that is, either for $p$ itself or its negation[24].

First, Past Indicative, that is, forms that signal direct evidence.

$$\begin{array}{ll} & [\![p\text{-}\mathbf{I}]\!] = \\ \text{Presupposition} & \mathit{DirectEvidence}_i(p) \vee \mathit{DirectEvidence}_i(\neg p) \\ \text{Assertion} & \qquad\qquad\qquad p \end{array}$$

where $\mathit{DirectEvidence}_A(p)$ means that agent $A$ has direct evidence that $p$ holds, and $i$ is the speaker. The *DirectEvidence* operator applies to formulas in the internal language (not propositions) and is most likely closed under the same inferential operations as belief ascriptions.

For Renarrated, I put the claim that there is some previous utterance (from some speaker $A$), from which $p$ has to be derivable, into the presupposition part, in order to make this claim immune to challenges, while still allowing it to contribute to the propositional content of the utterance. Among the proposals for dealing with the scope of evidentials, I am not sure the presuppositional one is the most preferable, but I chose it as the most familiar one. I indicate that the natural logic derivations $\mathbf{N}$ is the class of inferences allowed in deriving $p$. This class seems to be more or less the same as for deriving beliefs (see §4.7). However of the other operations used in deriving beliefs, internal perception cannot be based on another person's words, so it's inapplicable. Example (84) also shows that one cannot base the inference on the absence of certain information in a corpus of texts, so inference from ignorance is out of the question as well.

[24] von Fintel and Gillies (2009a) use a similar presupposition for the epistemic *must*; for them *must* $p$ presupposes that neither $p$ nor $\neg p$ is directly known.

At the same time, it is not a particular class of beliefs but the intentions of the original speaker that seems to count (see (83)), so this formal representation is still not completely adequate. The absence of direct evidence is implicated by the use of Renarrated, but not presupposed or asserted; the implication is cancelable, as witnessed by (106), the example where the speaker knows that hearsay information is false.

Nothing is asserted in a Renarrated utterance, but the hearer's attention is being drawn to the proposition $p$, so that the hearer can accept it or deny. I use a special kind of speech act here, Presentation, proposed by Faller (2002).

$$\begin{array}{ll} & [\![p\text{-}\mathbf{R}]\!] = \\ \text{Presupposition} & \exists A.\exists q.(\mathbf{say}(A,q) \wedge (\Box(\mathbf{B}_i q \to \langle \mathtt{N} \rangle \mathbf{B}_i p) \vee \Box(\mathbf{B}_i q \to \langle \mathtt{N} \rangle \mathbf{B}_i \neg p))) \\ \text{Presentation} & p \end{array}$$

Finally, for Conclusive I adopt the viewpoint of Stankov (1969): even though in most cases it signals inference as the source of belief, formally it serves as the least marked member of the opposition. $p$-**C** carries no presupposition and just asserts $p$. Conclusive is the only option for the cases where neither Indicative nor Renarrated are applicable. Most of such uses are those where there is some inference involved, thus the impression that Conclusive itself carries the inferential meaning.

It should be emphasized that the treatment I propose for hearsay evidentials is not an alternative to either the presuppositional theory of evidential contribution to meaning, to illocutionary operator theory or to the theory of not-at-issue assertion. In fact, my approach is compatible with any of these. I don't wish to take sides in the debate on how the evidential contribution fits into the meaning of the sentence as a whole; I'd like to specify as precisely as possible what that contribution consists in. My theory should be contrasted with the possible world account of propositions which serves as a basis both to (Izvorski, 1997) (in the form of Kratzer's theory of epistemic modals) and to the dynamic semantics of (Murray, 2010).[25] Any variant of possible-world based semantics will have problems with examples like (81) (going to Sofia) or (83) (seeing two bottles), where the required configuration of possible worlds is present, but the form of expression in the prejacent is too far from the original statement. Another advantage of my account is its ability to explain the choice of the inferential marker in examples like (92), where a mathematical calculation is involved. Also, as already mentioned in §5.2.6.1, since there is no requirement that hearsay information come from a single source, under the possible-world analysis anyone who has ever heard two contradicting statements should be justified to use Renarrated in stating any proposition at all.

On the other hand, my theory encounters problems with examples like Koev's (115) (Nixon's tapes), where the hearsay information (that Nixon had been involved in the destruction of the tapes) serves as a premise in the speaker's derivation of the prejacent, but the source of the hearsay information had no intention to inform the speaker about a particular event.

[25] In (Faller, 2002), the sincerity conditions of a hearsay evidential include the claim that $p$ has been asserted. There is no analysis as to what it takes to assert $p$ by uttering $q$, where $p$ and $q$ do not coincide.

Faller (2011) tries to clarify the meanings of evidentials using Kratzer's theory of modals, claiming that hearsay evidentials select their modal base and inferential markers make use of the ordering source.

## 5.5 Final remarks

Inferential evidentials in some languages are not used to indicate an arbitrary logical derivation; the inference class is restricted. One example is the Bagvalal indirect evidential (Tatevosov, 2007) or Cuzco Quechua *=chu-sina* (Faller, 2011) which can only signal the inference from the result of a process to the process itself.

Another is the Japanese *darou*, which can only be used for inferences from general regularities to a particular case (Hara, 2006). The mechanisms developed in this thesis, which allow us to specify classes of inferences, can be particularly useful in characterizing the meaning of such evidentials. So, if we use **G** as the name of the "generic inference" rule, we could analyze the Japanese $p$ *darou* as

$$\langle \mathrm{Infer}; \mathbf{G}; \mathrm{Infer} \rangle \mathbf{B} p$$

That is, the derivation of $p$ should include at least one use of the G rule.[26]

The alternative analysis for such limited inferentials is the modal one, using an appropriate modal base and ordering source. One class of examples that can distinguish between these two analyses is where the inference is non-defeasible — where there are no possible worlds that make the premises true and the conclusion false. In such cases evidentials should be allowed if they signal the type of inference, but superfluous if they are a kind of modals.

[26] It should not be possible to derive $p$ *without* the use of **G**. This may be considered a pragmatic restriction, or one could specify it explicitly:

$$[\![ p \ \mathit{darou} ]\!] = \langle \mathrm{Infer}; \mathbf{G}; \mathrm{Infer} \rangle \mathbf{B} p \wedge \neg \langle \mathrm{Infer}_{\setminus \mathbf{G}} \rangle \mathbf{B} p$$

where $\mathrm{Infer}_{\setminus \mathbf{G}}$ is the closure of the rules other than **G**.

Unfortunately, both inference from results and generic inference seem to be inherently defeasible rules, so finding such examples will be difficult.

# Chapter 6

# The Paradox of Clarity

## 6.1 The Problem, Barker and Taranto's solution

In Barker and Taranto (2003), Taranto (2006), Barker (2009), the construction *It is clear that p* is analyzed (as well as its variant *Clearly, p*).

As an initial approximation, the construction seems to mean that $p$ is entailed by the evidence available to some relevant group, which typically includes all the participants of the conversation. So, for example,

(117) It is clear that Abby is a doctor.

can be uttered when both the speaker and the hearer are looking at a picture of a woman wearing a lab coat and a stethoscope.

Barker and Taranto state the following problem: if the evidence presented to every participant of the conversation (part of the common ground) already entails $p$, there is no need in stating $p$. The common ground, viewed as a set of possible worlds, does not change after the assertion of clarity is made.[1]

The solution proposed by Barker and Taranto involves the notion of a "linguistic side effect". Every sentence is assigned some truth conditions,

[1] As (Barker, 2009) notes, there are cases where the set of relevant participants does *not* contain both the speaker and the audience. In these circumstances the paradox does not arise. One type of such cases are assertions of personal clarity.

and the dynamic effect of uttering it partly consists in narrowing the common ground by excluding those possible worlds that do not meet the truth conditions (this is the "main effect" of uttering the sentence). However, some changes to the common ground may not be related to the outside world, but to the state of the communication itself.[2] New discourse referents may be introduced. Standards may be set for vague predicates. For example,

(118) Bill is tall.

may be uttered to provide information about Bill's height (that is, for its main effect; this would be a descriptive use of *tall*), but it could also be uttered in order to specify what counts as tallness in the situation under discussion (a metalinguistic use of *tall*).

Clarity assertions, according to Barker and Taranto, are always used exclusively for their side effects. Namely, they set standards for what evidence is considered sufficient for belief in their argument proposition $p$ (Barker and Taranto, 2003; Taranto, 2006), or what evidence is considered appropriate justification for $p$ (Barker, 2009). That is, among the context elements constituting the common ground before the utterance, those are excluded where the standards of belief/justification in the current conversation are set too high.

The theory proposed by Barker and Taranto has it as its consequence that asserting the clarity of $p$ does not in fact entail $p$.

[2] Barker and Taranto (2003), following Stalnaker, note that the conversation itself is part of the world. For this reason, they do not consider it necessary to add additional information to the common ground apart from the set of possible worlds.

Barker (2009) treats the common ground as a set of pairs $\langle d, w\rangle$, where $d$ is a state of the conversation (including standards for vague predicates), and $w$ is a possible world. This is the formalism I use in discussing B&T's theory. The difference plays no role in what follows, however.

## 6.2 Problems with B&T

### 6.2.1 Factivity

There are, however, some problems with this theory. First, the prediction that *Clearly, p* does not entail $p$ is not borne out. This can be easily seen by considering cases where $p$ turns out to be false.

Considering cases where clarity assertions stand in the present tense, B&T can predict the infelicity of statements like

(119) #It is clear that Abby is a doctor, but in fact she is not.

The clarity assertion in the first clause ensures that the speaker believes Abby to be a doctor. But in this case she cannot sincerely utter the second clause, on pain of falling victim to Moore's paradox.

However, as soon as we put the example in the past tense, those pragmatic factors are no longer in play. *Clear* examples (120) pattern with simple statements of a proposition (121), not with expressions of belief (122) or justifiability assertions (123):

(120) a. #It was clear that Abby was a doctor, but in fact she was not.
b. It seemed clear that Abby was a doctor, but in fact she was not.

(121) a. #Abby was a doctor, but in fact she was not.
b. Abby seemed to be a doctor, but in fact she was not.

(122) a. We believed Abby to be a doctor, but in fact she was not.
b. It seemed to us that we believed Abby to be a doctor, but in fact she was not.

(123) a. It was justifiable to conclude that Abby is a doctor, but in fact she was not.

b. It seemed justifiable to conclude that Abby is a doctor, but in fact she was not.

In examples (120a) and (121a) we have a contradiction, which is absent in (122a) and (123a). In (120b) and (121b), the second clause denies correctness of the speaker's opinion expressed in the first clause. In (122b) and (123b) it does not.

Barker claims that *Clearly, p* patterns with belief assertions (non-factive) rather than knowledge assertions (factive), since they can be combined with *might not p* claims without contradiction. My intuitions differ from his on this point. Substituting an actual sentence for *p* in his examples to ease judgment, we get

(124) a. We know that Abby is a doctor, although she might not be.

b. We believe that Abby is a doctor, although she might not be.

c. It is clear that Abby is a doctor, although she might not be.

It seems to me that the only way to avoid inconsistency in (124c) is by making a pause after the first clause:

(125) It is clear that Abby is a doctor... wait a minute, she might not be one, she might be an actress.

One can repair hasty knowledge claims in a similar way. So clarity assertions pattern with factive statements after all.

### 6.2.2 Repeated clarity assertions

Secondly, in the Stalnakerian framework, once the standards of justification/belief are set, they can only get looser in the subsequent discourse (the context elements with tighter standards have already been eliminated). Consider, however, the following example.

(126) *A and B are sitting in an emergency room. A woman in a lab coat (X) walks along the corridor.*

A: This is clearly a doctor.

*A man (Y) walks by in the opposite direction. He wears a lab coat as well. He also has a stethoscope around his neck and carries a medical record under his arm.*

A: Clearly, this is another doctor.

Suppose we have four possible worlds:

$w_1$ X and Y are both doctors.
$w_2$ X is a doctor. Y is not.
$w_3$ Y is a doctor. X is not.
$w_4$ Neither is a doctor.

We also have three possible degrees of skepticism (these are part of the state of conversation; we are not interested in the other parts):

$d_1$ Wearing a lab coat is sufficient to be judged a doctor.
$d_2$ Wearing a lab coat is not sufficient, but together with a stethoscope and a medical record it does satisfy our doubts.
$d_3$ Nothing, even the medical record, is convincing enough.

Note that, since Y has more doctor-like features, there is no refinement of the vague standard for justification that would make X count as a doctor, but not Y.

At the start of the conversation, every world/standards combination is possible:

$$S_1 = \left\{ \begin{array}{l} \langle d_1, w_1\rangle, \langle d_1, w_2\rangle, \langle d_1, w_3\rangle, \langle d_1, w_4\rangle, \\ \langle d_2, w_1\rangle, \langle d_2, w_2\rangle, \langle d_2, w_3\rangle, \langle d_2, w_4\rangle \\ \langle d_3, w_1\rangle, \langle d_3, w_2\rangle, \langle d_3, w_3\rangle, \langle d_3, w_4\rangle \end{array} \right\}$$

After the first utterance, those world-standard pairs are eliminated that don't allow lab coat to count as enough evidence for doctorhood (note that the clarity assertion does not tell us anything about the world itself):

$$S_2 = \{\langle d_1, w_1\rangle, \langle d_1, w_2\rangle, \langle d_1, w_3\rangle, \langle d_1, w_4\rangle, \}$$

The second clarity assertion could serve to eliminate $d_3$ out of the set of possible standards, but these world-standard pairs are already eliminated by the time it is uttered. Thus, in Barker and Taranto's framework, the assertion would be uninformative and therefore infelicitous. However it is perfectly normal.

### 6.2.3 No vagueness

Contrary to Barker and Taranto's claim, clarity assertions can be used in situations where there is no vagueness at all and the standards for belief/justification are completely determined. In particular, mathematical discourse:

(127) Take an integer $n$ divisible by 9. Clearly, $n$ is also divisible by 3.

To accommodate these cases, B&T would have to argue either that there are vague standards of belief/justification involved (in particular, that there are context elements where $n$ is divisible by 9, but somehow not by 3), or that this kind of use is special and needs separate treatment. If they choose the latter option, an explanation would be in order, first, why the theory for mathematical (and similar) uses of *clear* does not apply to the more mundane situations, and second, why the polysemy of the *Clearly, p* construction is the same across a wide variety of languages.

## 6.3 Missing inference

My proposal is to take seriously the idea that the *clearly* construction marks the result of an inference. Namely,

(128) *It is clear to A from q that p* can be analyzed as: *A* has performed a sound inference which has *q* as premises and *p* as conclusion.

This is exactly what Barker (2009) calls the missing entailment theory (and rejects). On my analysis, *It is clear that p* does entail *p*. By asserting clarity, the speaker takes full responsibility for the soundness of her inference — even if the inference is defeasible. Thus, the behaviour of (120) is explained. In (126), the second utterance requires a separate (although similar) inference, so it is not superfluous.

Availability of clarity assertions for mathematical statements follows trivially on my account: these statements lose their special status; just like statements about the world, they are subject to inference operations.

If the *from q* part of the clarity assertion is omitted, then the evidence used as source for inference is left unspecified. In fact, when *A* is not the speaker, this inference may not be available to the speaker:

(129) I see that it is clear to you that John is lying; can you explain why?

If the *to A* part is left out, there should be an inference available to every participant of the conversation. Moreover, every participant should be able to make the same inference. This can be illustrated by the following example. Suppose John has read *Crime and Punishment*, Mary has read *The Brothers Karamazov*, and Peter *The Idiot*. When they gather to exchange their opinions, according to my intuition, it would not be appropriate for one of them to say:

(130) It is clear that Dostoevsky is a great novelist.

even though it is appropriate to utter

(131) It is clear to everyone here that Dostoevsky is a great novelist.

This kind of truth conditions requires that the speaker, in order to assert clarity of $p$, both be able to draw the inference herself and be able to attribute *the same* inference to the other participants. In order to attribute the inference to the other participants, she needs to know that they possess the premises $q$ of the inference. This, of course, still comes short of the definition of the common ground (for example, the other participants may not know that the speaker knows that they know $q$), but it becomes rather hard to construct the tests, and when they are constructed, it is hard to elicit definite judgments on the appropriateness of using *Clearly,* $p$ in such situations. So, for all practical purposes, my account predicts that the premises of the inference should be in the common ground when making clarity assertions without specifying the experiencer.

One way to capture the intuition that in a clarity assertion the speaker needs to have a specific inference in mind is to construct information states not just for individual agents, but for groups as well. There is a discussion in von Fintel and Gillies (2009b) of ways to aggregate information states. For their purposes, however, an aggregated state is one where all the information possessed by a group is pooled (that would correspond to an intersection of possible world sets or to set union of information states as sets of sentences). In order to obtain the common ground, one would need to take into account only the information every participant in the conversation has (thus, set union of possible world sets or set intersection of representations).

The requirement for the inference justifying the clarity assertions to be sound can be demonstrated by Gettier cases: suppose Abby is in fact a doctor, but she is dressed in a lab coat for a Halloween party, with a toy stethoscope around her neck. Under this scenario, the inference "lab coat means doctorhood" is not sound, and (117) is false.

Barker and Taranto's question 'why ever assert clarity?' receives a plausible explanation under this analysis: the speaker notifies the audience that the information they have ($q$) is sufficient to infer $p$. Each member of the audience is invited to build the inference for himself. The clarity statement can be used to build a greater confidence in the audience than simply stating $p$: upon deriving $p$, the hearer does not depend any longer on whether he trusts the speaker.

There are certain features noted by Barker and Taranto that any account of the *clearly* construction should be able to explain. Three of these features fall out immediately from my analysis. These are the inapplicability of clarity assertions to cases of direct evidence, information already explicitly stated in the conversation, and belief without proper justification (examples from Barker (2009)):

(132) $^{\#}$It is clear that Abby is wearing a stethoscope.

(133) A. Guess what? It turns out that Abby is a doctor!
B. $^{\#}$Now that you've told me this, it's clear that Abby is a doctor.

(134) $^{\#}$It is clear that God exists.

In all of these cases, there is no inference that allows us to assert *p*; therefore, the *clearly* construction is inappropriate.

## 6.4 Barker's objections

Reasons given in Barker (2009) for rejecting the "missing entailment" theory of the kind I am defending are the following. First, clarity assertions are often made when the proposition in question is not in fact entailed by the evidence:

(135) It is clear that Abby is a doctor

is said when she might in fact be an actress or dressed for a Halloween party. All we need to say is that the inference whose existence is stated by the clarity assertion may be defeasible: it can involve generic statements as premises or use other types of default rules. There will be no strict entailment in such cases.

Second, for some examples the missing entailment theory seems to predict wrong results:

(136) A. John is a bachelor.
B. $^{\#}$Clearly, then, he is not married.

(137) A. John ate a sandwich and drank a glass of beer.

B. $^{\#}$So it is clear that John ate a sandwich.

We can note that the inferences involved in these examples are extremely simple: subtyping in (136) and conjunction simplification in (137). So the missing entailment theory can be saved if we specify that the entailment in question should be substantial enough — not limited to certain easy types of inference.

The requirement that the inference be nontrivial may stem from the fact that people are reluctant to recognize certain simple inference steps as such.[3] After all,

(138) John ate a sandwich and drank a glass of beer. $^{?}$Therefore, he ate a sandwich.

does not sound all that natural either. If we do recognize absence of trivial inference as part of the meaning of *clearly*, it has to be a presupposition, as demonstrated by a negation test:

(139) A. John is a bachelor.

B. $^{\#}$It isn't clear that he is not married.

(only allowable if B is disputing A's claim).

The class of 'trivial' inferences may, for all I see, coincide with the class of inferences involved in ascribing beliefs to other persons:

(140) a. Bill believes that John is a bachelor. Therefore, he believes that John is unmarried.

[3] This explanation was suggested by Gennaro Chierchia (p. c.).

b. Bill believes that $n$ is divisible by 9. #Therefore, he believes that $n$ is divisible by 3.

In (140a), we have a valid inference. It is impossible for a competent speaker to believe someone to be a bachelor without believing him to be unmarried. The inference in (140b), on the other hand, is invalid, since humans are not logically omniscient.

Gradability, which Barker uses as another argument against the missing entailment theory, is discussed later, in §6.7.

## 6.5 *Clearly* vs. epistemic *must*

In von Fintel and Gillies (2009a), an argument similar to mine is made with respect to the epistemic *must*, and a similar solution is proposed:

> Epistemic modals signal that their prejacent is not directly settled by the salient kernel (where 'kernel' is a set of propositions — a structure that does not have the closure property — *G. B*).

However, *clearly* and *must* are not interchangeable.

- In the *clearly* construction, the existence of an appropriate inference is part of the assertion. Unlike *must*, *clearly* can take narrow scope with respect to operators like negation and tense.

(141) It is not clear to me that Abby is a doctor, but she might be.

(142) It was clear to me yesterday already that Abby is a doctor.

(This is a property that many epistemic modals have, but by no means all of them. For example, the English *have to* can be embedded under tense and negation operators.)

- *Must* does not have to be based on public evidence, even when the relevant group is not specified explicitly. In fact, there is no way to specify the relevant group in *must*.

- In certain situations, an inference can be marked by *must*, but not by *clearly*:

  (143) *John left two hours ago. Every participant in the conversation knows this.*

  a. John must be home by now.

  b. $^?$Clearly, John is home by now.

  Note, however, that once the premises of the inference are stated explicitly, clarity assertion becomes better:

  (144) John left two hours ago. It takes only half an hour to get home from here. Clearly, he is home by now.

  I can see two ways to explain this behaviour:

  1. Clarity assertions require the premises of the inference to be actively entertained by the participants of the conversation. In (143), the information required to deduce that John must have arrived by now sits in the background of the interlocutors' minds. Once it is foregrounded in (144), one can use the *Clearly, p* construction.

The assumption that inferences based solely on background information do not give rise to clarity seems to be refuted, though, by the following example:

(145) The economy is clearly in recession.

can be uttered “out of the blue”, without any preparatory foregrounding statements.

2. In (143), the “derivation” that leads to the conclusion involves operations that do not look very much like traditional inference steps: something like constructing mental scales and measuring distances on those scales. Presumably, such activity does not count as “easy” for the purposes of clarity assertion. On the other hand, in (144), the premises are stated linguistically, and the inference consists in a couple of standard rule applications.

- One can use *clearly* to signal an inference whose conclusion is already known to the speaker.

(146) Mary has been out of town for three days. She has not phoned. Clearly, I'm worried/$^{\#}$I must be worried.

What matters in this example is that an inference exists from public evidence that leads to the conclusion stated in the prejacent. Even though the speaker has more direct means of knowing the prejacent, the use of *clearly* is sanctioned.

As for the solution proposed by von Fintel and Gillies, it involves (as has been noted already) contexts as sets of propositions. Such a set induces a partition on possible worlds. Proposition $p$ is not settled by the context if there are possible worlds belonging to the same class in the partition which don't agree on $p$. Such a construction does not distinguish between equivalent propositions, so it is easy to build counterexamples to the theory using standard philosophical test cases:

(147) a. This animal has a heart.
      b. So, clearly, it has a kidney as well/So it must have a kidney as well.

(148) a. Triangle *ABC* is equilateral.
      b. Clearly, it is equiangular/So it must be equiangular.

Assuming that creatures with a heart are necessarily all and only creatures with a kidney, proposition *this animal has a heart* is exactly the same as proposition *this animal has a kidney.* Upon uttering (147a), this proposition is settled in the context of the conversation. (147b) (in either of its variants) should be abnormal. In fact, it is fine.

## 6.6 Special case: Sherlock Holmes

There is one special use case of clarity assertions, which can be demonstrated by an example, suggested by Derek Ball (p. c.). Imagine Sherlock Holmes investigating a case together with Dr. Watson and Inspector Lestrade. After all the evidence has been collected (and known by all participants in the conversation), Holmes points out the murderer:

(149) It is clear that the butler did it. For the maid was out on leave on the day of the murder, the gardener is deaf and would not hear the

doorbell. . . etc. etc.

What distinguishes this kind of usage from the standard one is that the inference that leads to the conclusion can be arbitrarily complex. After making the clarity assertion, the speaker immediately presents the inference.

Perhaps the simplicity of inference is just a pragmatic requirement in the standard clarity assertion cases. When the inference is not presented immediately by the speaker, the clarity assertion seems pointless as long as the audience is not able to recover it: the assertion does not increase the hearer's confidence in the proposition stated.

## 6.7 Gradability

As emphasized by Barker (2009), clarity is gradable: we have expressions like *crystal clear*, *somewhat clearer* etc. The theory presented here does not allow us to capture this property of clarity assertions. I have to resort to an informal description as to where the sources of gradability might be located.

There are several parameters by which inferences can be graded. Two are the length of inference and the likelihood of discovering it. As Barker's example:

(150) It is reasonably clear that Mars is barren of life.

shows, clarity is gradable with respect to the level of confidence that the inference provides to its conclusion. Most inferences in everyday life employ some amount of inductive and/or defeasible reasoning, so they don't guarantee the truth of their conclusion with absolute certainty. Moreover, people, with their limited reasoning capabilities, sometimes doubt whether the derivation they have just built qualifies as a valid (much less sound) inference. Conclusions of inferences that are really bulletproof can be characterized as *absolutely clear*, *crystal clear*, and inferences that employ a lot of heuristics, generic reasoning and such can give rise to statements about propositions that are *reasonably clear* or *relatively clear*. This analysis recovers much of the intuition behind Barker's theory. It also shows why mathematical inferences (even very long and complicated ones) hardly ever give rise to gradable clarity:

(151) $^{??}$It is reasonably clear that Fermat's Last theorem holds.

When sentences like (151) are used, this happens for the last of the reasons mentioned: the speaker does not have complete confidence that the proof she has in mind is correct.

The fact that clarity is gradable shows that we are unlikely to discover one day exactly what pattern of inference can give rise to clarity assertions: this is context dependent and vague. This vagueness can lead to side effects of the sort described by Barker: a clarity assertion may serve to establish which inferences count as easy. However, just like in cases with *tall*, such side effects do not exhaust the meaning of the construction.

## 6.8 Formal presentation

Two features distinguish clarity assertions from the other constructions studied in the thesis:

- There is a syntactic slot in the subcategorization frame of the construction for the premises of the inference:

  *It is clear to A* ***from*** *q that p*

- Inferences that give rise to clarity assertions are 'heavier' than those leading to indirect speech, belief assertions or hearsay evidentials.

The first point gives rise only to a technical difficulty. If the *from q* part is present, it needs to be reflected in the analysis of the construction. The one I propose is

$$\begin{array}{ll} & [\![\textit{It is clear to A from q that p}]\!] = \\ \text{Presupposition} & [\mathbf{L}_A](\neg\mathbf{B}_A q \to \neg\langle \mathit{Clear}_A\rangle \mathbf{B}_A p) \\ \text{Assertion} & \qquad\qquad p \wedge \langle \mathit{Clear}_A\rangle \mathbf{B}_A p \end{array}$$

where *Clear* is the class of 'easy' inferences (to be discussed in a minute) and **L** is a new elementary action — loss of a belief. The formula thus translates as "If $A$ loses the belief that $q$, he would not be able to easily deduce that $p$, but as it stands, $A$ can easily deduce that $p$."

There is still a number of details to be ironed out. We need to specify what happens when the non-obligatory syntactic arguments *from q* and *to A* are omitted. When $A$ is unspecified, it defaults to 'every participant in the current conversation' — a composite agent whose beliefs are the intersection of the participants' individual beliefs. When there is no explicit $q$,[4] it defaults to the *immediate* public evidence (certainly not the vast body of common world knowledge). This default works even if the $A$ argument is present. Consider

(152) It is clear to me that Peter is lying.

[4] Syntactically, $q$ cannot even be a finite sentence; it has to be nominalized. Thus we have

(1) It is clear by the look on Peter's face that he is lying.

or, at best, (proposition named, but not expressed)

(2) It is clear from the fact that Peter stutters that he lacks confidence.

I might have a much closer acquaintance with Peter than my audience, which allows me to recognize signs of dishonest behaviour, but the signs themselves should be public and they have to play a crucial role in my inference for (152) to be true.

Furthermore, the **B** operator in the formula above does not indicate simple presence of a formula in an agent's belief box, but 'established' presence (compare Slobin and Aksu-Koc (1986)). In fact, a clarity assertion with an explicitly specified agent $A$ indicates that $p$ is already present in $A$'s beliefs as the result of an inference. The 'belief loss' operator **L** subtracts established beliefs, and as a result the agent loses derived belief in $p$.

Another distinguishing feature of clarity assertions as compared with the constructions we studied so far is that 'easy' inferences for the purposes of this construction are much more complicated. In indirect speech and belief ascriptions we searched for a decidable class of inferences; this decidability allows the ascriber of an attitude to predict reliably that the ascribee would make the inference in question as soon as need arises. In the typical use of a clarity assertion — where $A$ is unspecified and the default includes both the speaker and the audience, the speaker does not believe that the audience has already made the inference. The clarity assertion itself serves to stimulate the audience to search for such an inference.

Inferences that serve in belief ascriptions and indirect speech — such as conjunction simplification and scalar inferences, — are typically performed without any conscious effort. They do not feel like inferences to a naïve speaker. In the case of clarity assertions, such a speaker is typically very much aware that a certain cognitive effort is required to reach the conclusion. Moreover, our class of admissible inference should be bound not only from the above (where the inference becomes too hard), but also from below (where it does not feel like inference anymore), as illustrated by examples like (136).

Unlike in belief reports or indirect speech, inferences employed in clarity assertions are very often defeasible. It is the presence of such defeasible rules that makes Barker and Taranto's theory work in a lot of use cases.

All of that said, I will not attempt to specify with any precision, the way I did in all the previous chapters, which class of inferences can justify clarity assertions. This class seems to be very much dependent on the context (something that is clear to a math professor may not be clear to a first grader). Moreover, whether a given inference falls into the class seems to be vague, leading to gradability of clarity assertions. A crude approximation could be to assign a weight to each type of derivation rule and limit a weighted sum of all rules used in a derivation.

# Conclusion

In Chapter 2 of this dissertation, a logical framework was introduced and investigated. In the subsequent chapters, we saw this framework applied to a number of natural language constructions. It makes sense to indicate how various features of the logical framework turned out to be useful in describing the linguistic data.

First of all, I have gained the greatest advantage from my decision to choose rule schemas as the elementary actions in my dynamic logic. On one hand, one needs to distinguish between applications of different rules; otherwise there is no hope to characterize various inference classes supported by constructions I consider. Thus, systems of Duc (2001) or Jago (2006) would be insufficient. On the other hand, if a single instance of a rule is treated as an elementary operation, in the style of Velázquez-Quesada (2011), one would not be able to name any such operation in an analysis of linguistic constructions. Quantification over elementary actions would be necessary and, whenever iterated applications of the same rule are permitted, over finite chains of elementary actions. In other words, rule schemas are the right level of granularity of actions when describing linguistic constructions.

On the other hand, only part of operations mentioned in the linguistic chapters of the dissertation can be specified using simple rule schemas of Chapter 2. This formalism is sufficient for Natural Logic and substitution of coreferential definite NPs. It is only partly sufficient for inference from ignorance (one has to supply a separate criterion that tells us where the rule is applicable). Similarly, defeasible inference steps, which can be used in clarity assertions, require complicated rules of applicability. Translation between languages, allowed in indirect speech, cannot be represented as a set of inference rules at all.

Iteration is required to express any class of inferences where a certain rule can be used an unlimited number of times. This is needed for indirect speech, belief ascriptions and the Bulgarian Renarrated. At the same time one cannot restrict inference classes acceptable in natural language constructions to those closed under iteration (that is, to classes where a set of rules used is limited, but not the number of instances of each rule or their order). In Chapter 4, I argued that existential generalization is not acceptable after “internal perception” operations in belief reports. Such an analysis explains the puzzle cases like the DENY problem, but as a result, belief inferences lack deductive closure in the sense of Konolige (1986). In the case of clarity assertions, whatever analysis one chooses in order to formalize the “easy inference” class employed in that construction, it should include a limitation on the number of inference steps, and thus will also not be expressible as an iteration closure over a set of rules.

It seems that the whole power of at least the language of Section 2.1 is needed to adequately describe the linguistic data.

# Bibliography

Aikhenvald, Alexandra Y.: 2004, *Evidentiality.* Oxford University Press.

Anand, Pranav and Andrew Ira Nevins: 2004, 'Shifty Operators in Changing Contexts', in R. Young (ed.), *Proceedings of Semantics and Linguistic Theory 14*, 20–37. CLC Publications, Ithaca.

Andrejčin, L.: 1944, *Osnovna bălgarska gramatika.* Sofia. (Basic Bulgarian grammar).

Andrejčin, L.: 1949, *Grammatika bolgarskogo jazyka.* Moscow. (A grammar of Bulgarian).

Artemov, Sergei N.: 1994, 'Logic of Proofs', *Annals of Pure and Applied Logic* **67**, 29–59.

Artemov, Sergei N.: 2004, 'Evidence-based common knowledge', Technical Report TR-2004018, CUNY Ph. D. Program in Computer Science.

Artemov, Sergei N. and Elena Nogina: 2005, 'Basic epistemic logic with justifications', Technical Report TR-2005004, CUNY Ph. D. Program in Computer Science.

Asher, Nicholas: 1986, 'Belief in Discourse Representation Theory', *Journal of Philosopical Logic* **15**, 127–189.

Atanassov, Dimka: 2010, 'The Bulgarian Reportative as a Conventional Implicature'. Poster at the Mid-Atlantic Colloquium of Studies in Meaning (MACSIM) workshop.

Audi, Robert: 1982, 'Believing and Affirming', *Mind* **91**, 115–120.

Bach, Emmon: 1986, 'Natural language metaphysics', in R. Barcan Marcus and P. Dorn, Gerald aand Weingartner (eds.), *Logic, Methodology and Philosophy of Science*, Vol. VII, 573–595. Elsevier.

Barker, Chris: 2009, 'Clarity and the grammar of skepticism', *Mind and Language* **24**, 253–273.

Barker, Chris and Gina Taranto: 2003, 'The paradox of asserting clarity', in P. Koskinen (ed.), *Proceedings of the Western Conference in Linguistics (WECOL) 2002*, Vol. 14, 10–21. Department of Linguistics, California State University, Fresno, CA.

Barnes, Janet: 1984, 'Evidentials in the Tuyuca verb', *International Journal of the American Linguistics* **50**, 255–271.

Beaver, David I. and Brady Z. Clark: 2008, *Sense and Sensitivity. How Focus Determines Meaning.* Wiley-Blackwell.

Bojadzhiev, T., D. Tilkov, S. Stojanov, and Popov K. (eds.): 1983, *Gramatika na săvremennija knizhoven bălgarkij ezik*, Vol. 2. Sofia. (A grammar of modern literary Bulgarian).

Boolos, George S., John P. Burgess, and Richard C. Jeffrey: 2002, *Computability and Logic.* Cambridge University Press.

Brasoveanu, Adrian and Donka F. Farkas: 2007, '*Say* Reports, Assertion Events and Meaning Dimensions', in A. C. et al (ed.), *Pitar Mos: A Building with a View. Papers in Honour of Alexandra Cornilescu.* Editura Universitatii din Bucuresti, Bucharest.

Cappelen, Herman and Ernie Lepore: 1997, 'On the Alleged Connection Between Indirect Speech and the Theory of Meaning', *Mind and Language* **12**, 278–296.

Carnap, Rudolf: 1947, *Meaning and Necessity.* The University of Chicago Press.

Chafe, Wallace and Johanna Nichols (eds.): 1986, *Evidentiality: the linguistic coding of epistemology.* Ablex, Norwood, NJ.

Crimmins, Mark: 1992, *Talk about Beliefs.* MIT Press.

Demina, E.I.: 1959, 'Pereskazyvatel'nye formy v sovremennom bolgarskom literaturnom jazyke', in *Voprosy grammatiki bolgarskogo literaturnogo jazyka.* Moscow. (Hearsay forms in modern literary Bulgarian).

Demina, E.: 1970, 'Kăm istorijata na modalnite kategorii na bălgarskija glagol', *Bălgarskij ezik* **XX**.

Dennett, Daniel: 1975, 'Brain Writing and Mind Reading', in K. Gunderson (ed.), *Language, Mind and Knowledge*, 403–415. Univ. of Minnesota Press, Minneapolis.

Ditmarsch, Hans van, Wiebe van der Hoek, and Barteld Kooi: 2007, *Dynamic Epistemic Logic*, 1st edn. Springer Publishing Company, Incorporated.

Duc, Ho Ngoc: 2001, *Resource-Bounded Reasoning about Knowledge*, Doctoral Dissertation, Univ. of Leipzig.

Elgot-Drapkin, Jennifer: 1988, *Step-Logic: Reasoning Situated in Time*, Doctoral Dissertation, University of Maryland.

Fagin, Ronald and Joseph Y. Halpern: 1988, ‘Belief, awareness, and limited reasoning’, *Artificial Intelligence* **34**, 39–76.

Fagin, Ronald, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi: 2003, *Reasoning about Knowledge.* MIT Press, Cambridge, Mass.

Faller, M.: 2002, *Semantics and pragmatics of evidentials in Cuzco Quechua*, Doctoral Dissertation, Stanford.

Faller, Martina: 2006, ‘Evidentiality below and above speech acts’.

Faller, Martina: 2011, ‘A Possible Worlds Semantics for Cuzco Quechua Evidentials’, in N. Li and D. Lutz (eds.), *Semantics and Linguistic Theory (SALT) 20*, 660–683. eLanguage.

Field, Hartry H.: 1978, ‘Mental Representation’, *Erkenntnis* **13**, 9–61.

Fielder, Grace E.: 1995, ‘Narrative Perspective and the Bulgarian *l*-participle’, *Slavic and East European Journal* **39**, 585–600.

von Fintel, K. and A. Gillies: 2009a, ‘Must…Stay…Strong…’, *http://mit.edu/fintel/fintel-gillies-2009-mss.pdf.*

von Fintel, Kai and Anthony Gillies: 2009b, ‘*Might* Made Right’, in A. Egan and B. Weatherson (eds.), *Epistemic Modality.* Oxford University Press, Oxford.

Fischer, John Martin: 1985, ‘Functionalism and Propositions’, *Philosophical Studies* **48**, 295–311.

Fitneva, S.: 2001, ‘Epistemic marking and reliablility judgements: evidence from Bulgarian’, *Journal of Pragmatics* **33**, 401–420.

Fitting, Melvin: 2005, 'The logic of proofs, semantically', *Annals of Pure and Applied Logic* 1–25.

Fodor, Jerry: 1975, *The language of thought.* Harvard University Press, Cambridge, Massachusetts.

Fodor, Janet Dean and Ivan A. Sag: 1982, 'Referential and Quantificational Indefinites', *Linguistics and Philosophy* **5**.

Friedman, V.A.: 1986, 'Evidentiality in the Balkans: Bulgarian, Macedonian, and Albanian', in W. Chafe and J. Nichols (eds.), *Evidentiality: the linguistic coding of epistemology*, 159–167. Ablex, Norwood, NJ.

Friedman, V.A.: 2000, 'Confirmative/nonconfirmative in Balkan Slavic, Balkan Romance and Albanian with additional observations on Turkish, Romani, Georgian and Lak', in L. Johanson and B. Utas (eds.), *Evidentials in Turkic, Iranian and Neighboring Languages*, 329–366. Mouton de Gruyter, Berlin.

Friedman, V.A.: 2002, 'Hunting the Elusive Evidential: The Third-Person Auxiliary as a Boojum in Bulgarian', in V. Friedman and D. Dyer (eds.), *Of All the Slavs My Favorites: In Honor of Howard I. Aronson* (*Indiana Slavic Studies* 12), 203–230.

Garrett, Edward John: 2001, *Evidentiality and Assertion in Tibetan*, Doctoral Dissertation, UCLA.

Groenendijk, Jeroen and Martin Stokhof: 1988, 'Type-shifting Rules and the Semantics of Interrogatives', in G. Chierchia, B. Partee, and R. Turner (eds.), *Properties, Types and Meaning. Vol. II: Semantic Issues*, 21–69. Reidel, Dordrecht.

Hara, Yurie: 2006, 'Non-propositional modal meaning'. ms., University of Delaware/University of Massachusetts, Amherst.

Harel, David: 1979, *First-Order Dynamic Logic.* Springer-Verlag, Berlin.

Horn, Lawrence R.: 1996, 'Exclusive Company: *Only* and the Dynamics of Vertical Inference', *Journal of Semantics* **13**, 1–40.

Hrakovskiy, V.S. (ed.): 2007, *Evidentsial'nost' v jazykah Evropy i Azii.* Nauka, Sankt-Peterburg.

Izvorski, R.: 1997, 'The Present Perfect as an epistemic modal', in A. Lawson and E. Cho (eds.), *Proceedings of SALT VII.* CLC Publications, Cornell University.

Jago, Mark: 2006, *Logics for Resource-Bounded Agents*, Doctoral Dissertation, Univ. of Nottingham.

Jago, Mark: 2009, 'Epistemic Logic for Rule-Based Agents', *J Log Lang Inf* **18**, 131–158.

Kamp, Hans: 1990, 'Prolegomena to a Structural Account of Belief and Other Attitudes', in C. A. Anderson and J. Owens (eds.), *Propositional Attitudes*, 27–90. CSLI.

Kamp, Hans, Uwe Reyle, and Josef van Genabith: 2005, 'Discourse Representation Theory', in D. Gabbay and F. Guenther (eds.), *Handbook of Philosophical Logic.*

Koev, Todor: 2010, 'Temporal Evidentiality as Not-at-issue Assertion'. draft, Rutgers university.

Konolige, Kurt: 1986, *A Deduction Model of Belief.* Pitman, London.

Konolige, Kurt: 1990, 'Explanatory Belief Ascription', in R. Parikh (ed.), *TARK9́0 Proceedings of the 3rd Conference on Theoretical Aspects of Reasoning and Knowledge*, 85–96. Morgan Kaufmann, San Francisco, CA.

Kratzer, A.: 1981, 'The notional category of modality', in H.-J. Eikemeyer and H. Rieser (eds.), *Words, worlds, and contexts*, 38–74. de Gruyter, Berlin.

Kripke, Saul A.: 1979, 'A Puzzle about Belief', in A. Margalit (ed.), *Meaning and Use*, 239–283. D. Reidel, Dordrecht.

Kutzarov, I.: 1984, *Preiskazvaneto v bălgarkija ezik.* Sofia. (Renarration in Bulgarian).

Kutzarov, I.: 1994, *Edno ekzotichno naklonenie na bălgarskija glagol.* Sofia. (On one exotic mood of the Bulgarian verb).

Lambek, Joachim: 1961, 'How to Program an Infinite Abacus', *Canadian Mathematical Bulletin* **4**, 295–302.

Lewis, David: 1970, 'General Semantics', *Synthese* **22**, 18–67.

Lycan, William: 1985, 'Tacit Belief', in R. J. Bogdan (ed.), *Belief: Form, Content and Function.* Clarendon, Oxford.

Maier, Emar: 2007, 'Quotation marks as monsters, or the other way around?', in M. Aloni, P. Dekker, and F. Roelofsen (eds.), *Proceedings of the Sixteenth Amsterdam Colloquium*, 145–150.

Maslov, Yu.S.: 1981, *Grammatika bolgarskogo jazyka.* Moscow. (A grammar of Bulgarian).

Matthews, Robert J.: 2007, *The Measure of Mind: Propositional Attitudes and Their Attribution.* Oxford University Press.

Matthewson, L., H. Davis, and H. Rullman: 2007, 'Evidentials as epistemic modals: evidence from St'át'imcets'. Ms., University of British Columbia.

McCready, Eric and Nicholas Asher: 2006, 'Modal subordination in Japanese: Dynamics and evidentiality', in A. Eilam, T. Scheffler, and J. Tauberer (eds.), *Penn. working papers in linguistics*, Vol. 12.1, 237–249. Penn Linguistics Club, University of Pennsylvania.

McCready, Eric and Norry Ogata: 2007, 'Evidentiality, Modality and Probability', *Linguistics and Philosophy* **30**, 147–206.

Mkrtychev, A: 1997, 'Models for the Logic of Proofs', in S. Adian and A. Nerode (eds.), *Logical Foundations of Computer Science '97, Yaroslavl'*, Vol. 1234, 266–275. Springer.

Moore, Joseph G.: 1999, 'Misdisquotation and Substitutivity: When Not to Infer Belief from Assent', *Mind* **108**, 335–365.

Moore, R.: 1985, 'Semantical Considerations on Nonmonotonic Logic', *Artificial Intelligence* **23**, 75–94.

Moore, R.C.: 1995, *Logic and Representation.* CSLI.

Moore, R.C and G.G Hendrix: 1979, 'Computational Models of Belief and the Semantics of Belief Sentences', Technical report, SRI International. Reprinted as Chapter 4 of Moore (1995).

Murray, Sarah E.: 2010, *Evidentiality and the structure of speech acts*, Doctoral Dissertation, Rutgers.

Nitzolova, R.: 2006, 'Vzaimodejstvie evidentsial'nosti i admirativnosti s kategorijami vremeni i litsa glagola v bolgarskom jazyke', *Voprosy jazykoznanija* 27–45. (Interaction of evindentiality and mirativity with the categories of tense and verb person in Bulgarian).

Nitzolova, R.: 2007, 'Modalizovannaja evidentsilal'naja sistema bolgarskogo jazyka', in V. Hrakovskiy (ed.), *Evidentsial'nost' v jazykah Evropy i Azii*, 107–196. Nauka, Sankt-Peterburg. (The modalized evidential system of Bulgarian).

Oswalt, Robert L.: 1986, 'The Evidential System of Kashaya', in W. Chafe and J. Nichols (eds.), *Evidentiality: the linguistic coding of epistemology*, 29–45. Ablex, Norwood, NJ.

Powers, Lawrence H.: 1978, 'Knowledge by Deduction', *Philosophical Review* **87**, 337–371.

Ågotnes, Thomas: 2004, *A logic of finite syntactic epistemic states*, Doctoral Dissertation, Department of Informatics, University of Bergen, Norway.

Rantala, Veikko: 1975, 'Urn models: a new kind of non-standard model for first-order logic', *Journal of Philosophical Logic* **4**, 455–474.

van der Sandt, Rob: 1992, 'Presupposition Projection as Anafora Resolution', *Journal of Semantics* **9**, 333–377.

Sauerland, Uli and Matthias Schenner: 2007, 'Embedded Evidentials in Bulgarian', in E. Puig-Waldmüller (ed.), *Proceedings of Sinn and Bedeutung 1*, 495–509. Barcelona.

Scatton, E.: 1984, *A reference grammar of modern Bulgarian.* Slavica Publishers, Columbus, Ohio.

Shan, Chung-chieh: 2010, ‘The character of quotation’, *Linguistics and Philosophy* **33**, 417–443.

Slobin, D.I. and A.A. Aksu-Koc: 1986, ‘A psychological account of development and use of evidentials in Turkish’, in W. Chafe and J. Nichols (eds.), *Evidentiality: the linguistic coding of epistemology*, 159–167. Ablex, Norwood, NJ.

Soames, Scott: 1989, ‘Direct Reference and Propositional Attitudes’, in J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan*, 393–419. Oxford University Press.

Stalnaker, Robert: 1984, *Inquiry.* MIT Press.

Stankov, V: 1969, *Bălgarskite glagolni vremena.* Sofia. (Bulgarian verbal tenses).

Stich, Stephen P.: 1983, *From Folk Psychology to Cognitive Science: the Case Against Belief.* Bradford Books/MIT Press, Cambridge, MA.

Stoichkova, S. and E Chausheva: 1995, ‘Preiskaznoto naklonenie v njakoi sredstvata za masovo osvedomjavane’, *Problemi na sociolingvistikata.*

Taranto, Gina: 2006, *Discourse Adjectives.* Routledge, NY.

Tatevosov, S.: 2007, ‘Evidentsial’nost’ i admirativ v bagvalinskom jazyke’, in V. Hrakovskiy (ed.), *Evidentsial’nost’ v jazykah Evropy i Azii*, 351–397. Nauka, Sankt-Peterburg. (Evidentiality and mirativity in Bagvalal).

Tzonev, V.: 1910, ‘Opredelenni i neopredelenni formi v” bălgarski ezik’, *Godišnik na Sofijskija universitet: Istorikofilologičeski fakultet* **7**, 1–18. (Definite and indefinite forms in Bulgarian).

Velázquez-Quesada, Fernando R.: 2011, *Small Steps in Dynamics of Information*, Doctoral Dissertation, Institute for Logic, Language and Computation (ILLC), Universiteit van Amsterdam (UvA), Amsterdam, The Netherlands. ILLC Dissertation series DS-2011-02.

Wansing, Heinrich: 1990, ‘A General Possible Worlds Framework for Reasoning about Knowledge and Belief’, *Studia Logica* **49**, 523–539.

Whitsey, Mark: 2003, ‘Logical Omniscience: A Survey’, Technical Report NOTTCS-WP-2003-2, School of Computer Science and IT, University of Nottingham.

Willett, Thomas: 1988, ‘A cross-linguistic survey of the grammaticalization of evidentiality’, *Studies in Language* **12**, 51–97.

Zamansky, A., N. Francez, and Winter. Y.: 2006, ‘A ‘Natural Logic’ inference system using the Lambek calculus’, *Journal of Logic, Language and Information* **15**, 273–295.