Philosophers' Imprint

VOLUME 12, NO. 14
SEPTEMBER 2012

# The Open Instruction Theory of Attitude Reports and the Pragmatics of Answers

Philipp Koralus

*Princeton University*



## 1. Introduction[1]

It is a quotidian though extremely remarkable fact that by assertively uttering sentences we can make claims about the world. Why do we take an utterance of a sentence to convey one claim rather than another? Our understanding of context and intentions plays an important role. But the most important factor arguably is our grasp of the *linguistic meaning* associated with the sentence. Some may think of this as the *semantic contribution* of the sentence. On standard theories in philosophy and linguistics, we understand a novel utterance in large part by deriving its linguistic meaning from the linguistic meanings of its parts and its grammatical structure.

According to one commonsensical view, the linguistic meaning of proper names is unproblematic: names just stand for things. The linguistic meaning of 'Otto' just is the individual Otto, etc. We could add that the linguistic meaning of a sentence like 'Otto smokes' just is the singular proposition that the individual Otto has the property of smoking. Letting '<*P*<*Y*>>' stand for the proposition that the property *P* applies to the individual *Y*, we can regiment this view as follows: The linguistic meaning of 'Otto smokes' is <smoke<Otto>>. This now classical view famously runs into trouble when we consider sentences that report on beliefs, desires, and other attitudes.

### *1.1 Complex interpretations and Kripke's Puzzle*

Imagine the scenario of Superman comics. Intuitively, an assertion of (1) in the scenario would be true, while an assertion of (2) would be

1. This paper has been long in the making, benefiting from discussions with several audiences, beginning in 2008 with the Philosophy Society, Australian National University, and followed by the Philosophy Program at the Graduate Center, City University of New York, and the Departments of Philosophy and Psychology at Princeton University. I am greatly indebted to Gideon Rosen, Gilbert Harman, Philip Johnson-Laird, and two anonymous referees. For helpful discussions and comments over the years, I also thank Angela Mendelovici, as well as Sarah-Jane Leslie, Paul Benacerraf, Delia Graff Fara, and Will Starr. I owe a special debt of gratitude to Jay Atlas, whose seminal work on semantic nonspecificity and neo-Gricean pragmatics provided the crucial starting point for the views I develop in this paper, and who provided insightful comments and criticisms through several drafts.

false. At the same time, 'Clark Kent' and 'Superman' stand for the same individual *s*, so purely singular propositions are not sufficient to distinguish what is intuitively conveyed by (1) and (2).

(1) Lois believes that Superman flies.

(2) Lois believes that Clark Kent flies.

An attractive response is to say that we must distinguish *what* is being represented and *how* it is being represented. (1) and (2) are partly interpreted as making different claims about the properties of certain representations that Lois has of *s*. It seems natural to say that we interpret (1) as conveying that Lois believes that *s* can fly, where Lois's relevant representation of *s* has the property of presenting *s* as a superhero. By contrast, we interpret (2) as conveying that that Lois believes that *s* can fly, where Lois's relevant representation of *s* has the property of presenting *s* as a nerdy reporter. Put differently, (1) and (2) are interpreted as making different claims about the type of guise under which Lois represents *s* as having certain properties.[2] The truth conditions of these interpretations are independent. This suggests the following regimentation of what (1) and (2) convey, where '*Superhero*' and '*NerdyReporter*' denote different properties of representations, to be distinguished from more ordinary properties:

(3) <belief<Lois, <fly<*s*, *Superhero*>>>>

(4) <belief<Lois, <fly<*s*, *NerdyReporter*>>>>

Schiffer has recently argued for a regimentation along those lines, originally due to Récanati.[3] I will call interpretations of this sort, including properties of representations, "complex interpretations".[4] For our purposes, complex interpretations correspond to propositions like (3) and (4) that include properties of representations, which I will call "PoRs" for short.[5]

In the above example, distinct properties of representations are associated with different names, but this is inessential. The same name can be associated with distinct properties on different occasions.[6] Imagine Peter does not know that Paderewski *the pianist* is the same person as Paderewski *the former Prime Minister of Poland*. Peter has heard Paderewski give a stunning recital, and as a result:

(5) Peter thinks that Paderewski is talented.

However, Peter does not think much of politicians, and hence:

(6) Peter does not think that Paderewski is talented.

Paradoxically, there is a true utterance of 'Peter does not think that Paderewski is talented' even though Peter believes that Paderewski is talented. The apparent paradox, due to Kripke, is resolved if one gives (5) an interpretation very roughly glossed as 'Peter thinks that Paderewski *represented in a pianist-way* is talented' and (6) one roughly glossed as 'Peter does not think that Paderewski *represented in a politician-way* is talented'.

*1.2 Default singular interpretations*

Not all attitude report interpretations seem to involve what I called complex interpretations that would include a characterization of how individuals are represented. Imagine three of your colleagues are named Nancy, Ruth, and Terence. Suppose Nancy reports,

(7) Terence believes that Ruth smokes.

2. Contrast a view like Crimmins' (1992), on which we specify *token* ways of representing, which Schiffer (2000) and others have criticized.

3. Schiffer (2000); Récanati (1993). In the interests of focusing the discussion, I am setting aside the question of whether we sometimes include properties of representations along with a property like smoking as well.

4. I adopt novel terminology here since experience has shown that existing terminology like "Fregean interpretations" and "MoP-laden interpretations" is likely to mislead when applied to the regimentation suggested here.

5. I do not use the more familiar term 'MoP' (Mode of Presentation), since it seems to have become strongly associated with the "token" view mentioned in footnote 2.

6. Kripke (1979)

Does the report convey information about *how* Terence represents Ruth? Intuitively, the answer is "No". Everyday attitude reports like (7) do not seem to be interpreted as conveying such information. What (7) conveys seems to be adequately represented by the singular proposition (8):

(8) <belief<Terence, <smoke<Ruth>>>>

These sorts of cases give initial plausibility to what is sometimes called the "naïve neo-Russellian" view that the semantic contribution of an attitude report sentence just is a singular proposition after all.[7] On such a view, a complex interpretation might be taken to result from a pragmatic implicature, added to a semantically determined purely singular interpretation. Few people now accept the view that attitude report sentences like (1) and (2) have the same truth conditions, barring "bribery, threats, hypnosis, or the like".[8] But then, why has naïve neo-Russellianism been so influential? I believe one important reason is that singular interpretations are, in fact, the *default* interpretation of attitude reports, as suggested by examples like (7).[9] What this means is that singular interpretations are chosen in the absence of a special reason to prompt the hearer to do otherwise. In my view, this explains why, to many people, including the present author, something about the naïve neo-Russellian view somehow "seems right", even after one meditates on cases like (5) and (6), etc.

Even names that are strongly associated with a certain way of representing an individual will not always lead to complex interpretations. Imagine detectives in a covert investigation so secret that witnesses are questioned under a ruse. Let's further imagine that they are discussing what various witnesses have said, to get a preliminary notion of who is likely to be the culprit. All believe that it is very unlikely that anybody believed to be of good character by a certain Ms. Smith is guilty. One detective asserts (9) to his colleagues.

(9) Smith believes that Suspect A is an upstanding citizen.

It seems to me that we do not interpret (9) as making any particular claim about *how* Smith represents Suspect A, let alone that she thinks of the suspect *as a suspect*. For (9), as in the everyday case exemplified by (7), the natural interpretation is singular.

Now, the view due to Schiffer discussed above takes the semantic contribution of attitude report sentences to be propositions including properties of representations, as in (3) and (4), where attitude report sentences are taken to be indexical with respect to those properties. In cases like (7) and (9), it seems more natural to say that no information is conveyed about ways of representing. As I see the issue, this creates what Schiffer calls the "meaning intention problem" for the indexical view: Often, if not most of the time, we do not take attitude report sentences to be uttered with any intention to convey information about ways of representing, nor do we have any inclination to distinguish ways of representing in our interpretation, but the presence of an indexical for properties of ways of representing makes it hard to respect that in interpretation.[10] If an attitude report sentence includes an indexical for a PoR, then something must be assigned to that indexical.

### *1.3 The task ahead*

I propose that it would be attractive to have a theory of proper names and attitude report sentences that allows us to say that the latter can be literally interpreted as conveying complex propositions like (3) and (4) as well as singular propositions like (8), depending on context.[11]

7. For example: Soames (1987); Braun and Saul (2002)

8. To use Richard's phrase. Richard (1990), p. 125. Soames himself has abandoned the view in Soames (2004).

9. Jaszczolt (2000) seems to share this view.

10. Schiffer (2000). As Schiffer makes clear, this problem is not directly about capturing *truth conditions* of intuitive interpretations.

11. I do not have the space to consider other proposals that begin with a different analysis of what is conveyed by attitude report sentences. For a prominent example, see Richard (1990), who includes expressions like 'Paderewski' in the *semantic contribution* of an attitude report and argues that context supplies

As noted, one proposal to make attitude report sentences relevantly context-sensitive is to say that they involve indexicality, though, as noted, this route makes it hard to capture both purely singular and complex interpretations.[12] Another option that has been explored by various authors would be to let attitude report sentences semantically encode singular propositions, which can be supplemented by pragmatic implicatures.[13]

Taking a more unusual starting point, I propose a view on which purely singular interpretations of attitude reports and what I called complex interpretations, involving a characterization of how individuals are represented, are both literal interpretations, while the linguistic meaning of attitude report sentences is neither ambiguous nor indexical with respect to them. Attitude report sentences are semantically *nonspecific* with respect to the interpretations discussed above. This puts me in at least partial agreement with Bach and with more recent works of Soames, who have recently argued that attitude report sentences with proper names in the that-clause are semantically nonspecific.[14] What seems to be missing so far is a clear theory of what exactly compositional semantics *does* specify on such a view and what pragmatic principles can yield the various interpretations in context. In the following, I will expound such a theory. I will then discuss some of its advantages.

## 2. The Open Instruction Theory of attitude reports

I propose to follow Chomsky in supposing that the *linguistic meaning* of an expression consists in a set of instructions to conceptual systems for constructing mental representations.[15] I add that the variety of interpretations of attitude reports discussed in the introduction reduces to different ways of executing the instructions that are the linguistic meaning of attitude report sentences. My account is called the "Open Instruction Theory" (OIT) because linguistic meaning leaves *open* which of these interpretations is the correct one. These sentences are just *silent* on the relevant differences in interpretation. In particular, the relevant differences leave *no trace* in the linguistic meaning of attitude report sentences and have nothing to do with syntax. The framework of linguistic meanings as instructions to create mental representations makes it possible to give clear content to the distinction that has been drawn by theorists like Atlas, Bach, and Soames between ambiguity, indexicality, and nonspecificity.[16] In the proposed framework, we can conceptualize ambiguity as a case of multiple distinct instructions that correspond to the same surface form, which would be independently lexicalized or correspond to different syntactic structures. We can conceptualize indexicality as a case of a set of instructions including a contextually filled-in variable. Finally, we can conceptualize nonspecificity with respect to a range of interpretations as a case of static instructions that leave open a range of possible ways of executing them, leading to a range of mental representations that corresponds to the range of interpretations.[17]

Of course, we do not intend attitude report utterances to assert *instructions*: we intend them to assert truth-evaluable claims. We simply do not assert the linguistic meanings of the sentences we use. Linguistic meaning is one particularly important part of what systematically accounts for our intuitive judgments of the "meaning" of sentences we encounter in everyday life. *Linguistic meaning* is not to be confused with the pre-theoretical notion of "meaning" that accompanies these judgments.

a translation manual that determines which expressions can be used to report on a person's belief. For some criticisms of this approach, see Soames (2002).

12. Schiffer (1977; 2000)
13. Soames (1987); McGlone (2007)
14. Soames (2004); Bach (2000)
15. Chomsky (1965; 2000)
16. Atlas (1977; 2005); Soames (2004); Bach (2000). This distinction is also discussed in detail in Koralus (2011; 2010).
17. Contrast theories postulating "weak" existentially quantified linguistic meanings or disjunctive linguistic meanings, criticized as inadequate analyses of nonspecificity by Atlas (2005).

OIT shares with Kamp's (1981) Discourse Representation Theory (DRT) and Heim's (1982) file change semantics (FCS) the view that literal utterance interpretation involves a level of representation beyond what is directly constrained by syntax; they are representationalist theories of interpretation. However, adopting a representationalist theory of interpretation does not yet guarantee that the nonspecificity of attitude report sentences will be analyzed in the right sort of way, as distinct from, for example, ambiguity or anaphora.[18] These distinctions have important empirical consequences that I will discuss in sections 5 and 6. Nor does adopting a representationalist theory guarantee that we can capture default interpretations of attitude report sentences as well as departures from these defaults. If we assign nonspecific linguistic meanings, we still need a pragmatic theory that works with inputs that do not have determinate truth conditions, which rules out classical Gricean proposals. As Levinson (2000) observes, the connection between pragmatics and representationalist theories of interpretation is under-theorized. This may be partly due to the fact that representationalist theories have received somewhat less attention in mainstream philosophy of language than I believe they merit.

Here is what may, in part, be responsible for this comparative lack of attention: many have worried that representationalist theories of interpretation are generally less attractive, because they can help themselves to linguistically unconstrained postulates. On the one hand, this raises the threat of having the system overgenerate interpretations. On the other hand, appeal to a level of representation that is not syntactically constrained might make proposals seem less explanatorily powerful, due to worries about unconstrained expressive power similar to those that motivate Chomsky's minimalist program in syntax, which attempts to do away with multiple levels of syntax-internal representations (Chomsky 1995; 2000).[19] I am highly sympathetic to the *spirit* of Stanley's (2002; 2005) principle insisting on linguistic constraint in semantics that these concerns motivate. However, I think this principle is ultimately too strong.[20] I believe that where we have reason to believe that linguistic meaning is silent, we just have to find other ways of constraining our theories beyond linguistic meaning. Since we characteristically reason with what is conveyed in conversation, it seems natural to look to the psychology of reasoning to motivate as many aspects of the Open Instruction Theory as possible. Even if we cannot immediately find independent support for every moving part of the theory, what counts is that we open ourselves to a new range of independent considerations, where syntactic theory no longer provides any.

According to Johnson-Laird's influential theory, people ordinarily reason by building *mental models* that stand in certain relationships to each other.[21] The theory is intended to explain our competence and performance in reasoning even with nonverbal inputs, and neuroimaging evidence is beginning to be available to show that the relevant operations are processed in areas of the brain that are language-independent.[22] Though reasoning and linguistic interpretation are closely related, the study of reasoning has its own subject, which allows us to look to the theory of reasoning for independent support in theorizing about semantics and pragmatics, just as we might look to syntactic theory for independent support. Nearly all aspects of my theory that go beyond what is syntactically determined will be described in terms

18. So-called "underspecification" in discourse representation theory, as discussed by Asher and Lascarides (2003), is taken to encompass lexical ambiguity and quantifier scope ambiguity, as well as anaphora. This notion of "underspecification", unlike that put forth by Atlas, Bach, and Soames, seems too broad. Atlas (1977; 2005), Koralus (2011; 2010), and Zwicky and Sadock (1975) discuss some of the empirical consequences of the distinction between nonspecificity, lexical and quantifier scope ambiguity, indexicality, and anaphora, such as different constraints on interpretation imposed by VP ellipsis, which I will turn to later.

19. These sorts of considerations create some appeal for formal reconstructions of the machinery of DRT in terms that do not require a fully independent level of representation of discourse, as in Groenendijk and Stokhof (1991) and Muskens (1996).

20. For a critique, see Récanati (2002).

21. Johnson-Laird (1983; 2008); Barrouillet et al. (2000); Byrne (2005)

22. Bauer and Johnson-Laird (1993); Knauff et al. (2001); Kroger et al. (2008)

of operations and representations independently required by mental model theory. For now, what I will take from mental model theory is that individuals are represented by individual "tokens" in a mental model, and properties and relations are represented by "features" bound to those tokens, where features and tokens are mental representations and where binding is a psychological relationship between mental representations. We will say that, for each proposition we can entertain, we can build a mental model that represents that proposition. Singular propositions decompose into individuals, properties, relations, and propositions. Mental models correspondingly decompose into mental model tokens, features, and other mental models. For example, Mary is represented by the mental model token MARY. The proposition <smoke<Mary>> is represented by the mental model SMOKE(MARY). The proposition <love<Mary, John>> is represented by the mental model LOVE(MARY, JOHN), etc.

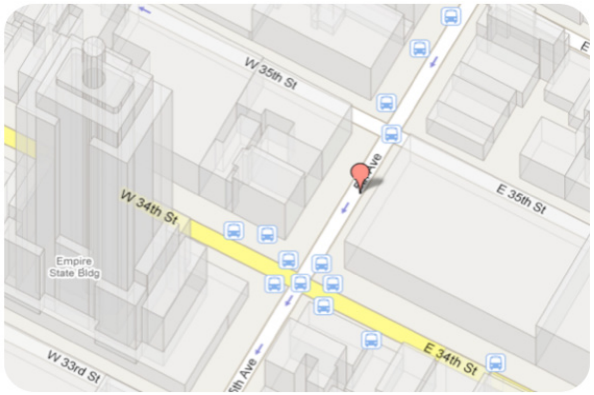
Mental model theory requires an operation that assembles mental models, which I will call the "model builder" (MB). On the Open Instruction Theory, utterance interpretation proceeds as follows: Given an utterance of a sentence *S*, the language faculty computes the *linguistic meaning* of *S*, which I will write as '‖*S*‖', not to be confused with the *denotation* of *S*. As noted, the linguistic meaning of *S* is a set of instructions to build mental models. The model builder then constructs a mental model, by executing ‖*S*‖. The resulting model expresses or represents a proposition. The claim we take to be made by an *assertive utterance* of a sentence *S* is the proposition represented by the mental model *M* that is obtained as the result of applying the model builder to ‖*S*‖. We judge the utterance of *S* to be true if and only if we judge the proposition represented by *M* to be true.

To flesh out this picture, we need an account of those mental models that amount to the interpretations discussed in the introduction. We then need recursive rules allowing us to compute the linguistic meaning of attitude report sentences and arrive at corresponding mental models. Finally, we need a pragmatic component of OIT that can make sense of both singular and complex interpretations.

*2.1 Capturing the interpretations of attitude report sentences*

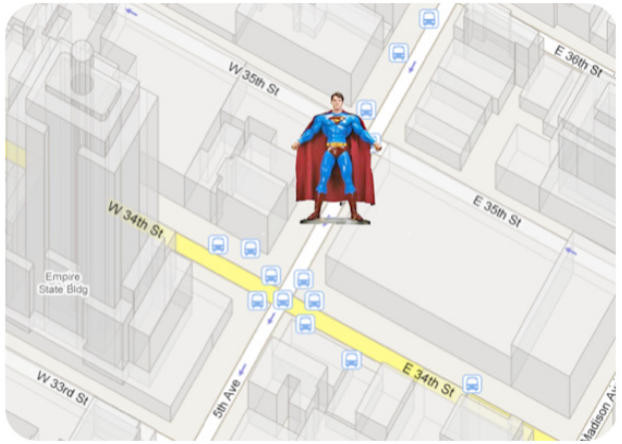
Suppose you are given an instruction to create a model representing Superman at a certain place. You could just fix a pin on an appropriate map, as in (10).

(10)




It is not conceptually mysterious that such instructions might leave open *how* individuals should be represented. You could instead use a detailed Superman figure that presents the individual in a more elaborate way:

(11)



You can represent an individual in your model with more or less elaborated placeholders. In addition to *what* is represented, there is a *way* in which it is represented. The instructions *left it open* how Superman was to be represented.

Just as instructions in the above toy analogy can leave open how an individual is to be represented, the instructions that amount to the linguistic meaning of an attitude report like (7) leave open how an individual, *e.g.*, Ruth, is to be represented.

(7) Terence believes that Ruth smokes.

The simplest way to represent an individual in a mental model is to use a mental model token that simply stands for the individual (like a simple pin in the toy analogy). For (7), this would give us the following mental model:

(12) BELIEF(TERENCE, SMOKE(RUTH))

We will take (12) to represent the singular proposition (13).

(13) <belief<Terence, <smoke<Ruth>>>>

Other mental models could serve as an interpretation of (7). Ruth might be like Paderewski, known by some as a famous pianist and by others as a famous politician. Perhaps Terence believes that only Ruth *represented in a pianist guise* smokes. In that case, our interpretive model might involve a representation of Ruth with the property of presenting Ruth as a pianist. I will take the relevant interpretive model to include a complex mental model token, comprising a mental model token representing Ruth bound to a mental model feature that represents *the property of presenting Ruth as a pianist*. I will write the symbol for a mental model feature of this sort as a superscript to the symbol for the mental model token to which it is bound. We obtain the following interpretive model:

(14) BELIEF(TERENCE, SMOKE(RUTH$^{\text{FAMOUSPIANIST}}$))

We take (14) to represent a proposition that includes a property of representations, viz. the property of presenting an individual as a famous pianist. For expository convenience, I will use combinations of capital letters in italics to denote properties of this sort.

(15) <belief<Terence, <smoke<Ruth, *FAMOUSPIANIST*>>>>

(15) is a proposition of the sort suggested by Récanati and Schiffer, discussed in the introduction, as corresponding to what I called a "complex interpretation". As noted, the truth conditions of a proposition like (15) are independent from a proposition like (16):

(16) <belief<Terence, <smoke<Ruth, *FAMOUSPOLITICIAN*>>>>

What about the relationship between (16) and (13)? It seems hard to avoid the view that (16) has to entail (13). Intuitively, we want (13) to be verified by any situation in which Terence represents Ruth as smoking and takes that representation to be true, regardless of how he represents Ruth. This has the consequence that (16) has to entail (13). This, in turn, quickly leads to the consequence that (17) is consistent and does not entail that Terence is irrational:

(17) <belief<Terence, <smoke<Ruth>>>>

& <belief<Terence, <¬smoke<Ruth>>>>

Similarly:

(18) <belief<Terence, <¬smoke<Ruth>>>>

$\nvdash$ <¬belief<Terence, <smoke<Ruth>>>>

Given the resources we have appealed to in order to account for various interpretations of attitude reports, the question arises as to what to make of a mental model like (19) that involves no representation of attitude relations and represents a proposition like (20).

(19) SMOKE(TERENCE$^{\text{FAMOUSPIANIST}}$)

(20) <smoke<Terence, *FAMOUSPIANIST*>>

Intuitively, it is truth-conditionally irrelevant *how* we represent Terence, if we want to ask if we are right about the world in representing that he smokes. (20) has the same truth conditions as (21).

(21) smoke<Terence>

This does not mean that there are no genuine differences in interpretation here. Suppose Terence also goes by the derogatory nickname 'Nitwit'. An interpretation of 'Nitwit smokes' may lead to a mental model in which Terence is represented in a derogatory way. Yet, intuitively, Nitwit smokes if and only if Terence smokes, because the property of smoking has nothing to do with *how* a smoking individual is represented. Ordinary properties and relations differ from attitude relations in this regard.

The next step will be to give an account of the linguistic meaning of attitude report sentences that allows us to generate the range of interpretations just discussed.

*2.2 Primitive instructions for proper names and verbs*

We need recursive rules that generate the linguistic meaning of attitude report sentences, based on primitive instructions. Where '‖*X*‖' denotes the linguistic meaning of *X*, and '*PN*' and '*V*' are variables for proper names and verbs:

(22) ‖*PN*‖ = "create a mental model token representing the bearer of *PN*".

‖*V*‖ = "create a mental model feature representing the property denoted by *V*".

(22) is an informal gloss. With the model-building operation MB, we can implicitly define instructions to create mental model constituents in terms of the result of applying MB to those instructions. Letting *t* be a mental model token, *j* a mental model feature, the first rule is:



(*R1*)

MB(‖*X*‖) ➔

- *τ, where τ represents i* — if *X* is a proper name of the individual *i*
- *φ, where φ represents P* — if *X* is a verb that denotes the property or relation *P*

The key to capturing the full range of attitude report interpretations was to leave open *how* individuals are represented in the interpretive model. We let *t*, above, range over both simple and complex mental model tokens. I will think of MB(‖*X*‖) as a mapping that takes us from the linguistic meaning of *X* to mental model constituents that are correct executions of it. For example, MB(‖Superman‖) can map ‖Superman‖ to any of the mental model tokens in (23). Which mental model token will, in fact, enter into the interpretive model will depend on pragmatic principles described in the next section. Some tokens are complex, in that they include features representing properties of representations:

(23)

| | | |
|---|---|---|
| MB(‖Superman‖) | SUPERMAN | *Simple token for singular interpretations* |
| | SUPERMAN$^{S}$ | *Complex token with feature S representing the PoR of presenting as a superhero* |
| | SUPERMAN$^{R}$ | *Complex token with feature R representing the PoR of presenting as a reporter* |
| | SUPERMAN$^{X}$ | *Complex token with feature X representing some other PoR* |

In formal terms, (*R1*) is a production rule that allows us to transition from MB(||*PN*||) to any available mental model token that represents the bearer of *PN* (more on constraints on the range of tokens in sections to follow).

### *2.3 Complex instructions and binding*

Mental model features bind to one or more mental model constituents. For example, a mental model feature SMOKE that represents the property of smoking binds to one mental model token representing an individual. A mental model in which SMOKE is bound to JOHN represents the proposition that John smokes. We should think of SMOKE as having an attachment site for an "agent" of the property represented by the feature. The feature representing smoking can be written as 'SMOKE(△)', where △ is a placeholder for a mental model token to which the feature SMOKE is bound.

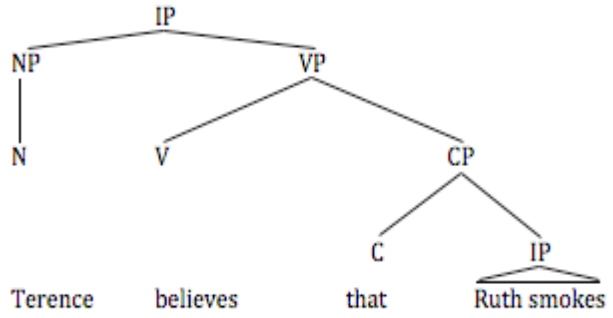
Mental model features representing relations, such as the mental model feature LOVE, can bind to two mental model tokens. Depending on how LOVE is bound to a mental model token, that token represents either the agent or the object of the love relation. Hence, the mental model feature LOVE(△, □), 'LOVE' for short, has two sites for attaching to mental model tokens: one for the lover and one for the beloved. LOVE is a two-place feature. I adopt the convention that the left-bound mental model constituent in the symbolic representation of a two-place mental model feature is the representation of the agent, and the right-bound constituent the representation of the object. Attitude features like BELIEF(△, □) bind to two mental model constituents, like LOVE. BELIEF binds to a (sub-)mental model *M*, as the object (the representation of what is being believed), as well as a mental model token $\tau$, as the agent (the representation of the believer). We will shortly have to specify how language controls how features get bound.

Now, the task is to give an account on which (7) determines an instruction to create a mental model that MB maps to mental models of the following two sorts: BELIEF(TERENCE, SMOKE(RUTH)), for a singular interpretation, or BELIEF(TERENCE, SMOKE(RUTH$^{F}$)), where F is a feature representing a property of representation, as needed for what I called "complex interpretations".

(7) Terence believes that Ruth smokes.

A simplified syntactic analysis of (7) will suffice for our purposes:

(24)




How do we let mental model features bind to the right mental model constituents? For simplicity, I assume a syntactic theory that marks argument roles in the syntactic structure. We need an operation that binds mental model features to mental model constituents in different argument roles. Call that binding operation '***B***'. For a mental model feature $\varphi$ and a mental model constituent □, where □ could be either a token $\tau$ or a (sub-)mental model *M*, ***B*** can, in principle, bind $\varphi$ to □ agent-wise or object-wise. Syntax determines whether binding is to be agent-wise or object-wise.[23] The binding operation ***B*** has two inputs: a set, {$\varphi$, □}, and an argument role parameter, $\theta$, which can

23. Another option would be to mathematically model mental model features as lambda abstractions, mimicking more classical approaches. We would get $\lambda\square.\lambda\tau$.BELIEVE($\tau$, □) for BELIEVE. On this option, mental model binding is modeled as function-application. Argument role marking in syntax would effectively be made redundant.

take the values "subject" or "complement", where the former specifies the agent role and the latter the object role. Let '__' stand for binding sites that ***B*** ignores.

(25)

$$\boldsymbol{B}(\theta,\{\varphi,\square\}) \rightarrow \begin{cases} \varphi\{\square\} & \text{if } \varphi \text{ is one-place and } \theta = \textit{subject} \\ \varphi\{\square,\_\} & \text{if } \varphi \text{ is two-place and } \theta = \textit{subject} \\ \varphi\{\_,\square\} & \text{if } \varphi \text{ is two-place and } \theta = \textit{complement} \end{cases}$$

Now, we do not have to explicitly formulate the instructions associated with complex expressions, as long as we can define the application of MB to complex instructions in terms of the primitives just discussed together with the binding operation.

The model builder MB applies to complex expressions as follows. With (*R4*), I assume, for simplicity, that 'that' is not directly contributing to the mental model.

(*R2*) MB(‖*IP*‖) ➔ ***B***(*θ*, {MB(‖*NP*‖), MB(‖*VP*‖)})

(*R3*) MB(‖*VP*‖) ➔ ***B***(*θ*, {MB(‖*V*‖), MB(‖*X*‖)}), $X \in \{NP, CP\}$ *for* $VP \neq V$

(*R4*) MB(‖*CP*‖) ➔ MB(‖*IP*‖)

Let's apply these rules to (7).

MB(‖*IP*‖) ➔ ***B***(*θ*, {MB(‖*NP*‖), MB(‖*VP*‖)}) [*R2*]

➔ ***B***(*subject*, {MB(‖Terence‖), MB(‖*VP*‖)})

MB(‖*VP*‖) ➔ ***B***(*θ*, {MB(‖*V*‖), MB(‖*CP*‖)}), [*R3*]

➔ ***B***(*complement*, {MB(‖believe‖), MB(‖*CP*‖)})



MB(‖*CP*‖) ➔ ***B***(*θ*, {MB(‖*NP*‖), MB(‖*VP*‖)}) [*R4*, *R2*]

➔ ***B***(*subject*, {MB(‖Ruth‖), MB(‖smokes‖)})

So,

MB(‖*IP*‖) ➔ ***B***(*subject*,{MB(‖Terence‖),***B***(*complement*,{MB(‖believe‖),***B***(*subject*,{MB(‖Ruth‖), MB(‖smokes‖)})})})

MB(‖*IP*‖) ➔ ***B***(*subject*, {MB(‖Terence‖), ***B***(*complement*, {BELIEF, ***B***(*subject*, {MB(‖Ruth‖), SMOKE})})})

By (*R1*), we can insert either a simple or a complex mental model token in place of 'MB(‖Ruth‖)'. For a simple token, we get:

MB(‖*IP*‖) ➔ BELIEF(TERENCE, SMOKE(RUTH))

For a complex token with a PoR feature F, we get:

MB(‖*IP*‖) ➔ BELIEF(TERENCE, SMOKE(RUTH$^{F}$))

Hence, (*R1*)–(*R4*) allow us to generate both singular and complex interpretations of attitude report sentences. Of course, the rules allow us to generate even more interpretations, such as (26), with some PoR feature X for Terence:

(26) BELIEF(TERENCE$^{X}$, SMOKE(RUTH))

Even though adding PoR features to TERENCE does not lead to different truth conditions, as discussed, we will eventually need an explanation of why we ordinarily don't get such interpretations in practice (setting aside cases like 'Nitwit believes that Ruth smokes', discussed above).

Some may find it objectionable that on this account of the linguistic meaning of attitude report sentences, as many interpretations are possible as the hearer can, in principle, distinguish properties of ways of representing. But this is how it should be! As Bach has convincingly argued, for any two such distinguishable properties, we can construct

a Paderewski-style scenario that requires distinguishing them to make sense of an attitude report.[24] The question of why we do not, in practice, have to consider an arbitrarily large set of possible interpretations to make sense of an attitude report will be answered by the pragmatic component of my theory, supplied below.

Taking stock, both singular and complex interpretations of attitude report sentences are accommodated by the proposed characterization of linguistic meaning, without appeal to ambiguity or indexicality. On the above proposal, attitude report sentences are indeed nonspecific with respect to the interpretations at issue. The distinction between these interpretations is not marked in the grammar. At the same time, we have given a precise characterization of how the nonspecific linguistic meaning of attitude report sentences with proper names in the that-clause is systematically generated from syntax and the lexicon.[25]

What remains is to give an account of how we settle on a particular mental model from the range of mental models that would count as correct executions of the instructions that are the linguistic meaning of attitude report sentences. This theory has to explain why singular interpretations are the default. It also has to explain when the model builder departs from the default interpretation.

### *2.4 From inference to the best interpretation to inference to the most responsive interpretation*

I will adopt the following insight due to Atlas and Levinson: We interpret an utterance by tacitly making an inference to the best interpretation.[26] Here, inference to the best *interpretation* is roughly analogous to inference to the best *explanation* in science.[27] Unfortunately, Atlas and Levinson's particular interpretive principles seem to me to be inapplicable to attitude report sentences and to create difficulties I would like to avoid. Atlas and Levinson reduce "inference to the best interpretation" to "inference to the most informative interpretation".[28] More specifically, they propose that the information content of a given interpretation is to be understood in some cases as the set of its potential falsifiers, following Popper, or in other cases as its set of logical consequences, following Carnap. The proposal has consequences that are somewhat puzzling, in my view.[29]

I suggest that we should save Atlas and Levinson's crucial insight that pragmatic principles bridging the gap between linguistic meaning and particular interpretations amount to principles of inference to the best interpretation and look for a new set of pragmatic principles that can be applied to attitude report sentences in the semantic framework I proposed. The key idea behind inference to the best *explanation* is that we begin with a set of hypotheses compatible with the data.[30] We would then choose the best hypothesis, "based on considerations such as which hypothesis is simpler, which is more plausible, which explains more, which is less *ad hoc*, and so forth".[31] Now, on the proposal of linguistic meaning developed in the previous section, a sentence

24. Bach (2000)

25. By contrast, from Soames (2004) and Bach's (2000) discussions, it is not clear how to provide a compositional account of the linguistic meaning of attitude report sentences that reflects their nonspecificity.

26. Atlas and Levinson (1981); Atlas (2005); Levinson (2000)

27. The term "inference to the best explanation" is due to Harman (1965).

28. Atlas and Levinson (1981); Atlas (2005)

29. The proposal is operationalized via the idea that normalized statements fronted by more existential quantifiers are less informative by the Popperian measure. Yet empirical existential statements with non-finite domains of discourse just do not have (finite) falsifiers. It is puzzling why adding more existential quantifiers would, in any sense, further decrease informativeness. Moreover, consider: '(A) There is a pink elephant in your room.' '(B) There is a pink elephant in your room, and there is a blue rhinoceros in your car.' Regimenting the predicates with some harmless simplifications, we get: '(A') ∃x(Ex & Px).' '(B')∃x∃y((Ex & Px) & (Ry & Cy)).' By the Popperian existential quantifier measure, (B) is *less* informative than (A), even though (B) entails (A). This seems odd. Any falsifier of (A) is also a falsifier of (B), so it just cannot be the case that (B) has fewer falsifiers. Atlas (p.c.) holds that we should say that (A) is intuitively more informative because it is more "specific". I prefer to avoid the complications just sketched.

30. Harman (1965) contrasts this with enumerative induction, where one would *deduce* a hypothesis from a list of observations via an induction principle.

31. Harman (1965)

encodes instructions that allow various mental models to be generated. We will say that the set of these mental models corresponds to what would be the set of alternative hypotheses for inference to the best explanation. The best *interpretation* then corresponds to the best mental model. To obtain substantive predictions, we need a proposal about the factors that determine which mental model is the best. I suggest that one factor is simplicity, in the sense of the number of mental model constituents. More will be needed.[32] It is a familiar move in theoretical debates to say of a set of competing theories (none falsified by extant data) that the best theory would answer such-and-such questions when combined with our background beliefs. This may be part of what the IBE theorist wants to capture with the notion of a certain theory being more explanatory than others. I suggest that literal utterance interpretation involves something similar, proceeding *relative to a background question*. One might plausibly suppose that, by default, the background question is something like, "What are some facts about the environment I don't know yet?", which is narrowed to something more specific in almost any actual dialogue and may be shifted to, *e.g.*, "What are some facts I don't know yet that hold in the world of the *Iliad*?", say, when we listen to a story. I will call the virtue of an interpretation of answering relatively more questions "responsiveness". I propose that, for a hearer *H*, inputs to literal interpretation are a sentence, background beliefs of *H*, and questions *H* would like to answer by engaging in discourse. Moreover, I propose to define the principles of inference to the best interpretation using resources independently required by the mental model theory of reasoning.[33] To avoid confusion with Atlas and Levinson's proposals that do not assign a special role to questions, I will call the proposal "Inference to the Most Responsive Interpretation" (IMRI):

(*IMRI*) Given an utterance of a sentence *S*, a mental model of background beliefs *BG*, and a mental model of background questions *Q*, the model builder will produce the mental model *M* that represents the best interpretation of *S*. *M* is the mental model that satisfies the following constraints:[34]

(*Literality*) Let the set of mental models that can be generated by executing ‖S‖ be $I_S$. *M* is a member of $I_S$.

(*Responsiveness*) If a model $M_1$ conjoined with the mental model *BG* (representing background beliefs) represents more answers to questions (represented by the mental model *Q*) that the hearer wants to answer by engaging in discourse than $M_2$ similarly combined, then $M_1$ is more responsive than $M_2$ relative to *BG* and *Q* (see (E1)–(E3) below). Let the set of all mental models in $I_S$ such that no other mental model in $I_S$ is more responsive when conjoined with *Q* and *BG* be $J_{BG,S}$. *M* is a member of $J_{BG,S}$.

(*Simplicity*) Let $K_{BG,S}$ be the set of mental models in $J_{BG,S}$ such that, for each model conjoined with *BG*, the resulting model is such that no other mental model in $J_{BG,S}$ conjoined with *BG* has fewer constituents.[35] *M* is a member of $K_{BG,S}$.

I will assume that BG is made up of the same sorts of constituents as interpretive models. We now need a working account of how background questions are represented in mental models, and we need

32. Needless to say, it is a separate question whether inference to the best explanation is the normatively correct account of how we should draw conclusions about theories in science. To the extent that we have reason to believe that it is correct, there is the potential to argue that the proposed view of utterance interpretation is a *rational* model in a fairly strong sense. Compare the Gricean project of accounting for implicatures in terms of a rational theory of communication.

33. Johnson-Laird (2008)

34. For now, I will postpone a discussion of what happens when there is no unique mental model. Moreover, I will narrow my view exclusively to *literal* utterance interpretation.

35. We will count constituents in our regimented representation of mental models the way we count free variables. For example, 'x loves x and wishes x to succeed' has one free variable; 'x loves y' has two free variables. For example, on this way of counting, the mental model LOVE(JOHN, JOHN) has one constituent less than LOVE(JOHN, MARY).

an account of how mental models are conjoined. I will provisionally adopt the view that a background-question mental model has the set of its exclusive alternative answers as its representational content.[36] A background question is then represented in the mental model framework as a set of mental models, each representing an alternative answer to the question.[37] Johnson-Laird's theory of reasoning with mental models independently requires mental models consisting of further mental models representing alternatives.[38] I will use '$[M_1/M_2/\ldots/M_n]$' to stand for a mental model that consists of mental models $M_1$ $M_2$, …, $M_n$, where each mental model has an alternative proposition as its representational content.

I have encountered two objections to the view just presented. One objection centers on an alleged vicious circularity. The worry is that the proposed theory of interpretation presupposes a question that itself needs to be interpreted. But, of course, the mental model representing our background question does not have to enter our cognitive economy by verbal means, any more than our background beliefs have to come to us through a verbal briefing before the beginning of a conversation. Another objection is that it does not seem intuitively *impossible* to approach an utterance without a particular question in mind. Nothing in (IMRI) forces us to deny this intuition. If we have no background question—if we are in a state that we can think of as an empty question model—no interpretation can (with an important caveat[39]) be more or less responsive than its alternatives, with the consequence that $J_{BG,S} = I_S$. Finally, to forestall a possible misunderstanding, note that the principles of (IMRI) apply only to literal utterance interpretation; they do not, in their present form, yield predictions about more general "conversational implicatures" that are not *literal* interpretations. By design, (IMRI) only generates predictions about how sentences are interpreted relative to a particular analysis of their linguistic meaning, which has to leave open a range of literal interpretations in order for the predictions to be nontrivial. (IMRI) is concerned with obtaining literal interpretations in the face of semantic nonspecificity. Classical Gricean pragmatics begins only once we have a literal interpretation, and it could, in principle, be combined with (IMRI). Moving beyond Grice, a reviewer alerted me to the need to distinguish (IMRI) from Hobbs et al.'s theory of "interpretation as abduction".[40] On their view, "the process of interpreting sentences in discourse can be viewed as the process of providing the best explanation of why the sentences would be true".[41] On their theory, the input to their pragmatic process is a semantically determined logical form in "first-order predicate calculus".[42] The pragmatic component then produces an "elaborated logical form", which entails the semantically determined logical form. The added material amounts to the explanation of *why* the semantically determined logical form is true. By contrast, on my account, the input to the pragmatic component does not have determinate truth conditions, which a statement in predicate calculus would have, merely determining a range of possible interpretations. The result of pragmatic principles is an interpretation with truth conditions. In addition, my view, unlike Hobbs et al.'s theory, is defined in terms of responsiveness to a background question that the hearer seeks to answer. Finally, abductive principles that would yield an explanation of *why* an utterance is

36. Note that, at least in principle, the view about *background questions* for the purposes of (IMRI) does not commit one to the view that this is the right account of the semantics of interrogative sentences in natural language. See Mascarenhas (2009). See Hamblin (1973) for the original notion of questions denoting sets of alternatives.

37. Since this yields mental models with infinitely many compartments for a question like "What is your favorite number?", we will need to say something about the possibility of representing alternatives implicitly in mental models, which Johnson-Laird's (2008) theory of reasoning independently requires. More on this below.

38. Johnson-Laird (2008)

39. There is, in fact, a way in which we could usefully think of an interpretation as less responsive than its alternatives, even in the *absence* of a background question: the interpretation could somehow raise a question.

40. Hobbs et al. (1993)

41. Hobbs et al. (1993), p. 69

42. Ibid., p. 75

true are far more powerful than what I propose. In my view, Asher and Lascarides correctly criticize Hobbs et al. for failing to distinguish literal utterance interpretation from a general account of how we update our beliefs.[43]

Moving on, we will also need a mental model constituent that represents negation, which is also already a component of the mental model theory of deduction.[44] I will use the symbol '¬' to denote the mental model constituent representing negation, when it occurs in the context of mental models, and to denote negation, when it occurs in the context of propositions.

Next, we need an account of how mental models are conjoined. Johnson-Laird is committed to a certain procedure governing how mental models are conjoined. This procedure can be regimented formally, though a sketch will suffice for the purposes of this paper, with more technical detail given elsewhere.[45]

So far, we have talked only about mental models representing simple propositions. For what follows, we will need more structure. We will say that mental models are sets of mental model features bound to mental model tokens. So, where we previously just talked about a mental model SMOKE(JOHN) as representing the proposition that John smokes, we will now talk about that mental model as a set {SMOKE(JOHN)} that has, as its unique element, the feature SMOKE bound to the token JOHN. To represent both the proposition that John smokes and the proposition that Mary smokes in a single mental model, we take {SMOKE(JOHN), SMOKE(MARY)}, which has two elements, etc. We will say that {SMOKE(JOHN), SMOKE(MARY)} has {SMOKE(JOHN)} as a sub-model. As in previous sections, if applications of mental model operations are not directly at issue and we are dealing with mental models that have unique sub-models, I will suppress set notation in exposition.

43. Asher and Lascarides (2003)

44. Johnson-Laird (2008)

45. Koralus (*in preparation*)



Besides mental models with multiple distinct sub-models, we will need mental models of alternatives. We will take a mental model of alternatives to be a set of sets—namely a set of mental models, which are, in turn, sets of mental model features bound to mental model tokens. For ease of exposition, we will use '[' and ']' instead of '{' and '}' and '/' instead of ',' to represent sets of mental models. For example, '[{GUILTY(SUSPECTA), ¬BELIEF(SMITH, UPSTANDING(SUSPECTA))} /{¬GUILTY(SUSPECTA)}]' denotes the set of alternative mental models comprising {GUILTY(SUSPECTA), ¬BELIEF(SMITH, UPSTANDING(SUSPECTA))} and {¬GUILTY(SUSPECTA)}. More generally, if $M_1$ and $M_2$ are mental models, then $[M_1 / M_2]$ is a mental model representing the alternatives represented by $M_1$ and $M_2$. Finally, we will add to our inventory a special primitive "null" mental model $M_\varnothing$ that represents contradiction.

We can now give a simplified definition of Johnson-Laird's main procedure for conjoining mental models, which appears to work as follows:

> (*Conjoin*) Given two mental models $M_1$ and $M_2$, **CONJOIN**$(M_1 , M_2) \rightarrow M_X$.
>
> **If** $M_2$ is a model of alternatives $[M_2/M_3/\ldots]$, then
>
> $M_X = [$**CONJOIN**$(M_1 , M_2)/$**CONJOIN**$(M_1 , M_3)/\ldots]$.
>
> **If** $M_2$ is drawn from a set of alternative mental models $[M_2/M_3/\ldots]$ and a sub-model $M$ of $M_1$ is also a sub-model of at least one of the alternative models in $[M_2/M_3/\ldots]$ but not a sub-model of $M_2$, then $M_X = M_\varnothing$.
>
> **If** $M_1 , M_2$ are mental models and there is an $M$ such that one of $M_1 , M_2$ has $M$ as a sub-model and the other has $\neg M$ as a sub-model, then $M_X = M_\varnothing$.

**If** $M_1$ , $M_2$ are mental models and one of $M_1$ , $M_2 = M_\varnothing$, then $M_X = M_\varnothing$.

**Otherwise**, $M_X = M_1 \cup M_2$.

A full account of the mental model theory of reasoning requires further principles, and some revisions may be necessary. Yet the above suffices to describe the core of (IMRI) in terms of mental model procedures. Background-question answering corresponds to conjoining question models and models representing putative answers. We can then define principles governing relative responsiveness as follows:

(E1) An interpretation $I$ is *maximally* responsive to $Q$ if **CONJOIN**((**CONJOIN**($I$, $BG$)), $Q$) yields a mental model in which the alternatives in $Q$ have been reduced to one (*i.e.*, no occurrences of '/' in the regimentation), distinct from $M_\varnothing$. If the one remaining alternative is $M_\varnothing$, we say that $I$ rejects the question.[46]

It may be hard for an interpretation to be *less responsive* than an interpretation that leaves the background question as it is or that rejects the question.[47]

(E2) An interpretation $I$ has zero responsiveness to $Q$ if **CONJOIN**((**CONJOIN**($I$, $BG$)), $Q$) yields a mental model in which no alternatives in $Q$ have been eliminated (*i.e.*, no reduction in occurrences of '/' in the regimentation).

(E3) For two interpretations $I_1$ and $I_2$ that rule out some but not all alternatives when conjoined with $BG$ and $Q$, $I_1$ is more responsive to $Q$ than $I_2$ iff more (explicit[48]) alternative models are ruled out by **CONJOIN**((**CONJOIN**($I_1$, $BG$)), $Q$) than by **CONJOIN**((**CONJOIN**($I_2$, $BG$)), $Q$) (*i.e.*, one eliminates more occurrences of '/' in the regimentation than the other).

Both mental model conjoining and the ability to compare numbers of alternatives are cognitive resources that are independently required by the mental model theory of reasoning.[49] Thus, the proposed theory of inference to the best interpretation can be defined in terms of cognitive resources that are independently motivated. It must be noted that the number of mental models that can be entertained at a time is finite. Yet potentially open-ended questions like 'Who will come to the party?' may correspond to an unbounded number of alternatives that have to be represented by mental models. We can only *implicitly* represent the mental models that represent those alternatives. The mental model

46. Suitably analyzed, this might be observed in cases like 'A: Whom did John kiss? B: John did not kiss anyone.'

47. For the purposes of this paper, I'm setting aside the possibility of an interpretation generating additional questions. However, we could add a principle (E4) along the following lines: An interpretation $I$ is negatively responsive ("inquisitive") if conjoining $I$ with $BG$ and $Q$ yields a mental model with *more* alternatives than in $Q$. Such a notion may capture the intuitive sense that some interpretations "raise more answers than they provide". It should be explored whether there are language-independent reasons to believe that highly controversial or unlikely propositions tend to independently raise questions in being combined with background beliefs and whether this accounts for why we may prefer, all other things being equal, uncontroversial and likely interpretations.

48. Though I have no space for a formal treatment in this paper, once we add Johnson-Laird's distinction between explicit and implicitly represented mental models, we can account for the fact that, even given background questions with infinitely many alternatives, some incomplete answers seem better than others because they may eliminate more *explicitly represented* alternatives. For example, suppose that someone who knows that Mike has a girlfriend asks him, 'Who will come to your party?' If Mike is an ordinary person, it would seem much more responsive for Mike to say, 'Well, my girlfriend won't come' than to say, 'Well, Bill Clinton won't come.' The account argued for by Johnson-Laird, Legrenzi, et al. (1999) suggests that ordinary reasoners tend to represent the relative probability of an event via the relative number of explicit alternative mental models that represent the event as occurring. If this is correct, then a partial answer that rules out an event we considered relatively likely will be relatively more responsive by (E3), because it eliminates more explicitly represented alternatives. There is no need to stipulate an independent preference for information-theoretically more informative interpretations.

49. Johnson-Laird (2008); Johnson-Laird et al. (1999)

theory of reasoning itself fundamentally requires a precise account of implicit mental models, but extending my formal treatment to capture this fact is, as a reviewer has convinced me, beyond the scope of this paper and is to be discussed elsewhere.[50] We can now consider what predictions the theory makes about the interpretation of simple sentences as well as about attitude report sentences.

*2.5 IMRI and the interpretation of attitude report sentences*

On the view presented so far, syntax and lexical items generate instructions to create mental models. Depending on the sentence, these instructions can allow for a range of mental models. Given this range of mental models, (IMRI) determines the mental model that will, in fact, be generated and whose representational content corresponds to the intuitive content of the utterance. The instructions encoded by (7) (see derivation at the end of section 2.3) allow for interpretive mental models of the following sort:

(7) Terence believes that Ruth smokes.

(27) BELIEF(TERENCE, SMOKE(RUTH))

(28) BELIEF(TERENCE$^{X}$, SMOKE(RUTH))

…

(29) BELIEF(TERENCE, SMOKE(RUTH$^{Y}$))

…

(30) BELIEF(TERENCE$^{X}$, SMOKE(RUTH$^{Y}$))

…

The mental models of particular concern are (27), corresponding to a singular interpretation, and interpretations of the form of (29) corresponding to complex interpretations involving PoR features of the sort involved in 'Superman' and 'Paderewski' cases. (28) and (30) correspond to interpretations with some PoR feature on mental model tokens outside of the attitude context. As noted earlier, interpretations of this sort are possible but nonstandard.

*2.5.1 Default interpretations in unembedded sentences*

Consider an utterance of (31) with no relevant background beliefs and a default background question along the lines of "What are some facts about the world I don't know about?"

(31) John smokes.

As far as OIT semantics is concerned, MB could generate SMOKE(JOHN) or any number of models like SMOKE(JOHN$^{POR}$) including some PoR feature or other. As defined earlier, the propositions represented by those models are true in all the same circumstances, corresponding to the intuition that the truth conditions of literal interpretations of (31) are basically context-invariant. As I suggested, by default, the background question is about (non-mental) facts about the world. This means that, given a default background question, the literal interpretations (in the sense of (IMRI)) of ||John smokes|| would tend to rule out the same set of alternatives. If JOHN$^{POR}$ is recognized as a correct execution of ||John||, then it must be a background belief of the hearer that any ordinary feature applying to JOHN$^{POR}$ also applies to JOHN and vice versa. As a result, SMOKE(JOHN$^{POR}$) combined with background beliefs would, at most, be as responsive in the sense of (IMRI) as SMOKE(JOHN). However, SMOKE(JOHN) is the simplest model, so (IMRI) correctly predicts that a PoR-free singular interpretation is the default interpretation of (31).

*2.5.2 Default interpretations with attitude verbs*

Why are we generally so good at making sense of other people? On a very plausible view, we have considerable success at figuring out other people's mental lives because we tend to attribute our own background beliefs and simple inference patterns to the people we are trying to understand, except for those beliefs and inference patterns we

50. Koralus (*in preparation*)

have special reason to think are not shared.[51] One way to follow this strategy would be to explicitly represent that we and other people represent alike. However, this seems needlessly complicated. Unless we have a special reason to think that our way of representing diverges from others' in respects that make a difference for our purposes, it seems easiest simply not to mark any differences. Let's take a case with a default background question (*e.g.*, "What are some facts I don't know yet?") and a background belief model that includes only the information that Nancy, Ruth, and Terence are colleagues of the hearer. Now, Nancy utters,

(7) Terence believes that Ruth smokes.

If the simplest way for us to represent the fact that Ruth smokes is to entertain the model SMOKE(RUTH), then unless we have a special reason to contemplate the possibility that Terence represents this fact in some different way, we can just interpret (7) as BELIEF(TERENCE, SMOKE(RUTH)). If we assume that we do not distinguish our way of representing from that of other people unless we have a special reason to do so, and we assume that we normally represent facts in the simplest way available to us, then no complex interpretation in a case like that of (7) is going to be better than a simple interpretation.

Now, let's return to the example involving detectives in a covert investigation so secret that witnesses are questioned under a ruse. Again, imagine that they are discussing what various witnesses have said, to get a preliminary notion of who is likely to be guilty. All believe that it is very unlikely that anybody believed to be of good character by a certain Smith is guilty.[52] One detective asserts (9) to his colleagues.

(9) Smith believes that Suspect A is an upstanding citizen.

51. Nichols and Stich (2003), p. 65

52. The fact that Smith may not arrive at the right judgment if, say, Jack the Ripper were introduced to her in a sheep costume, and innumerable other possible things that could cause a breakdown of Smith as a reliable judge (drugs, sleep deprivation, etc.), may simply be folded into a general measure of uncertainty of Smith's reliability in practice.

In the introduction, I noted that we do *not* intuitively interpret (9) as making any particular claim about how Smith represents Suspect A, let alone that she thinks of the suspect *as a suspect*. For (9), the natural interpretation is simple.

The background question is whether Suspect A is guilty. Conjoining this question with the background belief that Suspect A's being guilty is (simplifying a bit) incompatible with Smith believing of that individual that he is upstanding, we get a mental model of alternatives of the following sort:

[{GUILTY(SUSPECTA),¬BELIEF(SMITH, UPSTANDING (SUSPECTA))} / {¬GUILTY(SUSPECTA)}]

An interpretive model of the following sort would then be maximally responsive:

(32) BELIEF(SMITH, UPSTANDING(SUSPECTA))

The key idea is that if our background beliefs and background question do not already involve distinctions of ways of representing an individual that bear on how the alternatives in a question can be reduced, an interpretive mental model of an attitude report that includes PoR features cannot be more responsive than a simple default interpretation. Unless our background beliefs conjoined with our background question independently lead to a set of alternatives where some of those alternatives are distinguished by different properties of ways of representing, then no PoR-laden interpretation is going to be more responsive than a simple one. On the assumption that the alternatives determined by background questions and background beliefs normally do not hinge on distinguishing different PoRs, because, by default, we tend to be interested in what the environment is like and what facts about the environment are represented by people, (IMRI) predicts that simple interpretations of attitude report sentences, corresponding to singular propositions, are the default. One upshot is that processing the linguistic meaning of an attitude report sentence

together with (IMRI) never requires considering more different PoR features than are already included in the background question conjoined with background beliefs, so the set of alternative interpretations to consider is naturally limited.[53]

*2.5.3 Non-default interpretations involving attitude verbs*

Suppose Terence tells his colleague Nancy,

(33) John believes that Ruth smokes. John does not believe that Ruth smokes.

Intuitively, we would expect these utterances to give rise to a reaction like "Hey, wait a minute! What is going on here?" Faced with (33) without special background information, it is hard to figure out how to proceed with interpretation.

On the proposed theory, this is not surprising. Suppose we give a default interpretation to the first sentence in (33). We get the interpretive model:

BELIEF(JOHN, SMOKE(RUTH))

Suppose this then becomes part of our background belief model *BG* for the purposes of interpreting the next utterance. If we now were to give the second sentence a default interpretation as well, we would get the following interpretive model:

¬BELIEF(JOHN, SMOKE(RUTH))

Now, to assess the responsiveness according to (IMRI) of the second interpretation relative to a background question $Q$ and background beliefs *BG*, we have to look at the number of alternatives in

**CONJOIN**({¬BELIEF(JOHN, SMOKE(RUTH)), BELIEF (JOHN, SMOKE(RUTH)), …}, $Q$)

This reduces to **CONJOIN**($M_\varnothing$, $Q$) and further reduces to $[M_\varnothing]$, regardless of the content of $Q$. By (E1), this means that this interpretation amounts to rejecting the background question, changing the subject. The conversation cannot proceed in a normal incremental fashion.

If we instead use an interpretive model for the second sentence that includes a PoR feature, then the collapse into $M_\varnothing$ is blocked. However, nothing in context tells us what sort of PoR feature we should use. As a result, neither staying with a default interpretation nor departing from it in the absence of further contextual information makes it possible to proceed normally. This seems to correspond well to the intuitive reaction that would be produced by (33) in the absence of a special context.

In contrast to the case just discussed, the Paderewski examples, as normally described, include contextual information that suggests what sort of PoR features we should include in our interpretations. Peter does not know that the famous pianist by the name of Paderewski and the politician by the name of Paderewski are the same individual. Peter follows an announcement that Paderewski is playing the piano on the radio and comes to think that Paderewski's performance is rather good. Someone reports:

(34) Peter believes that Paderewski is talented.

In this case, we assume that our mental model of Peter's beliefs already includes two mental model tokens with different PoR features, where both of those tokens represent Paderewski. One of the tokens has a "pianist" PoR feature, while the other has a "politician" PoR feature. The question, then, is which of those tokens will be included in interpretation, in case this becomes necessary for a responsive interpretation. It seems to me that any successful theory would have to invoke a notion of a relative degree of salience of information established in the interpreter's background in such cases. On the present theory, this would mean a relative degree of salience of different mental model

53. This is not to say that the theory presented so far requires that people actually consider a full range of explicitly represented alternative interpretations. (IMRI) can be seen as a competence theory of utterance interpretation in a Chomskyan sense.

constituents.[54] We would say that, within the background belief model, some mental model tokens are more salient than others and that the best interpretation will use a mental model token whose occurrences already in *BG* are more salient than alternatives. This avenue does not seem altogether unattractive. This means we need to add the following principle to (IMRI) from section 2.4:

(*Salience*) If there is no unique member of $K_{BG,S}$, then $M$ is the model whose constituents are the most salient to the hearer at the moment of interpretation.

Finally, consider a case in which we know from context that there are multiple ways of representing an individual but where the local context does not tell us which to use for a particular utterance. Suppose somebody reports,

(35) Lois believes that Superman is talented. Lois does not believe that Clark Kent is talented.

As before, without adding a PoR feature to our interpretation of (35), we will again be rejecting the question, as in the case of (33), discussed above. However, our background belief model of Lois's beliefs will include at least two sorts of mental model tokens for the individual Superman: one with a "superhero" PoR-feature and one with a "reporter" PoR feature. However, nothing in context is telling us which of these tokens should be preferred.

Obviously, what breaks the symmetry are the names. It should be fairly uncontroversial that proper names are associated with various properties, in a way that nobody would take to be part of their linguistic meaning. These associations plausibly influence what information is salient to us. Presumably, the more a proper name occurs in the company of certain descriptive information, the more that information becomes associated with that name. This may explain why choosing a name that is shared by well-known people with socially desirable traits rather than undesirable ones is considered important by many people who are choosing a name for a child. There is a widespread intuition that common associations with a name influence what traits are perceived as salient, particularly in the absence of better knowledge of the person.[55] To account for the example at hand, we say that 'Superman' makes a "superhero" PoR feature more salient, while 'Clark Kent' makes a "reporter" PoR feature more salient.

In the rest of this paper, I will briefly discuss some of the relative advantages of the Open Instruction Theory.

### 3. Modeling other minds and interpreting attitude reports

Intuitively, (36) feels at least as problematic as (33):

(36) John believes that Ruth smokes. John believes that Ruth does not smoke.

This intuition is interesting, since it would be hard to argue that (37) is *inconsistent*:

(37) <belief<John, <smoke<Ruth>>>> & <belief<John, <¬smoke<Ruth>>>>

Believing *of Ruth* that she smokes seems indisputably compatible with believing *of Ruth* that she doesn't smoke. One may simply not realize that one's beliefs are of the same person, so (37) can be true of John even if John isn't irrational. If this is correct, then we should be able to have default interpretations for the two sentences in (36) without issue. But then, why would (36) seem at least as bad as (33)?

54. This effectively means adopting an idea that was suggested by Lewis, who wrote that to make sense of how certain expressions are interpreted, it may be necessary "to appeal to a salience ranking not of individuals but rather of individuals-in-guises" (Lewis 1979).

55. One advice column notes, "You want to avoid the baggage attached to infamous people or places" (http://baby-name-generator.com/BabyNaming4.html, accessed on1/13/11). A best-selling book on baby names notes, "On the first day in school, the teacher does a roll call. The only thing the other students know about a child is what he or she looks like and what his or her name is. Kids are likely to form a quick opinion from just those facts" (Lansky 1999).

I propose that what explains the fact that (33), without a special context, seems bad has to do with a further background assumption about other people's beliefs that we naturally make in interpreting multiple attitude reports. There seem to be good empirical reasons to think that when we process information about a person's beliefs, we use this information to build an integrated representation of that person's beliefs, and it seems very plausible that we ordinarily use the same reasoning procedures applied to our own model of the world to gain insight into the beliefs of other people.[56] In the framework of OIT, I propose that this is reflected in a background assumption that we can consolidate our mental models of other people's belief contents in the same way in which we can conjoin the mental models that correspond to our own beliefs. This would amount to the assumption that the below consolidation rule is freely applicable.

(E4) **CONSOLIDATE**(BELIEF($T$, $M_1$), BELIEF($T$, $M_2$)) ➔ BELIEF ($T$, **CONJOIN**($M_1$, $M_2$)

With this consolidation assumption, if we give the second sentence in (37) a default interpretation, after a default interpretation of the first sentence has already passed into our background beliefs, we get

BELIEF(JOHN,**CONJOIN**({¬SMOKE(RUTH)}, {SMOKE(RUTH)}))

which reduces to

BELIEF(JOHN, $M_\varnothing$)

This means that, given our background assumption of belief consolidation, (36) uttered in the absence of a special context seems to attribute irrationality to John. Together with a further plausible background assumption that people are minimally rational, we can explain the effect that such an utterance of (36) seems like it can't be quite right. It might have seemed, at first, that building an integrated representation of someone's beliefs based on successive belief reports, just as we might integrate our own mental models, is innocuous. However, on a closer look, it amounts to a substantive further assumption. In general, the belief relation is not closed under **CONJOIN**.

In the case at hand, belief consolidation together with the nature of mental models guarantees that further belief reports can't genuinely add to our model of John's beliefs, leaving us stuck with BELIEF(JOHN, $M_\varnothing$). To a certain extent, there no longer remains a substantive question as to what John believes once we admit BELIEF(JOHN, $M_\varnothing$), so it practically would not make sense to continue the discourse if the background question is concerned with John's beliefs. If we do believe that there is a substantive question of what John believes, then we may have among our background beliefs ¬BELIEF(JOHN, $M_\varnothing$). In this case, default interpretations of (37) would lead to a rejection of the question, in the sense of (E1) and (*Responsiveness*).

If we instead use interpretive models for (37) that include PoR features, then the collapse into $M_\varnothing$ is blocked. However, as in earlier examples, nothing in context tells us what sort of PoR feature we should use. The principles of inference to the best interpretation cannot settle on a unique best interpretation. As a result, it appears that the oddness of (37) out of context is explained. A default interpretation would basically force us to change the subject of conversation, while there is no determinate non-default interpretation that could avoid this. In sum, one advantage of OIT is that it can be readily integrated with independently motivated views about how we make sense of other minds, which can help explain certain intuitions that are puzzling if we exclusively focus on propositional content.

## 4. Interpretations that diverge from recognized communicative intention without appeal to charity

(IMRI) predicts that the best interpretation of certain attitude reports in context is such that the hearer knows that that particular interpretation is *not intended* by the speaker. On this point, (IMRI) diverges from

56. Nichols and Stich's (2003) discussion of our mentalizing capabilities contains many observations that support this view.

a more traditional view of utterance interpretation according to which the hearer, as a receiver, tries to recover the message that the speaker, as the sender, intends to transmit.

You and I know that not everybody knows that Paderewski-*the-statesman* is the same person as Paderewski-*the-pianist*. John and Mary are two people who are ignorant in this way. Imagine John sees Mary handing Paderewski a cigarette. He thinks he is looking at Paderewski-the-pianist. John comments to us,

(38) Mary believes that Paderewski smokes.

Later, John sees a similar scene. About to give a speech as a statesman, Paderewski is offered a cigarette by Mary. This time, Paderewski refuses. John thinks he is seeing Paderewski-the-politician and does not realize that he is seeing the same person as before. Mary similarly remains ignorant. John comments to us,

(39) Mary believes that Paderewski does not smoke.

In the context described, we intuitively interpret (38) as conveying a belief about Paderewski represented in one sort of way and interpret (39) as conveying a different belief about Paderewski represented in another sort of way. However, we cannot obtain these interpretations if utterance interpretation consists in recovering the communicative intention of the person making the utterance, as I will now argue.

As noted before, interpretations of attitude reports that convey PoRs are confined to cases in which there are special features in the background that suggest to us that we have to distinguish different sorts of representations of the same individual. Yet, from John's perspective, (38) and (39) require no more special efforts to make fine-grained distinctions about representations than (40) and (41) require for us:

(40) Mary believes that [*Elizabeth*] Harman is at the department reception.

(41) Mary believes that [*Gilbert*] Harman is not at the department reception.

From the perspective of John, there should be no intention to convey a PoR. From his perspective, there is no more need for a PoR to distinguish what he takes to be two Paderewskis than there is a need for PoR to distinguish the two Harmans from our perspective. If John *ever* has an intention to convey PoR-free, singular attitude reports, he should have that intention when he utters (38) and (39).

We naturally interpret these sentences in a way that includes PoRs to compartmentalize Mary's beliefs about Paderewski-the-pianist and her beliefs about Paderewski-the-statesman. But again, this interpretation cannot plausibly be what John *intends* to convey. An advantage of the view I have proposed is that the intuitive interpretation is predicted by the same pragmatic rules that would predict interpretations in ordinary cases (the case here is parallel to the case in the previous section). Unless the background question of the hearer specifically is concerned with how the *speaker* represents the world, our knowledge of what the speaker does or does not likely intend to communicate does not influence how (IMRI) determines the best literal interpretation.[57] From the perspective of (IMRI), the example discussed in this section is just a further manifestation of the fact that our default background questions are primarily aimed at the environment. By contrast, Asher, who discusses a similar example, has to say that "default rules can be overridden by a constraint [he calls] 'charity'".[58] On my account, the intuitive interpretations are the result of a direct application of the rules, not an exception.

It is very plausible that we do often employ some kind of charity principle in interpretation. However, to deal with the case described, quite a strong principle would be needed. Recall that <belief<*i*, <*P*<*x*>>>> & <belief<*i*, <¬*P*<*x*>>>> is not a contradiction, nor does it entail an attribution of irrationality. In fact, the relevant instantiation of this pattern in the example is true! This means that, for example, if we were to prefer a weak principle of charity along the lines of

57. Throughout, we have been concerned exclusively with literal interpretation, not with a general class of conversational implicatures.

58. Asher (1986), p. 146

"maximize the truth or rationality in the subject's sayings",[59] we would have no reason to diverge from what we take the speaker to have had in mind in the example at issue. (IMRI) allows us to avoid having to postulate a very strong principle of charity just to deal with the case under consideration.

On the view I have proposed, the pragmatic principles that are involved in arriving at *literal* utterance interpretations are not directly concerned with speaker intention, and making an appeal to a principle of "charity" is unnecessary to deal with the example discussed.[60] The speaker's primary opportunity at making the hearer consider a particular proposition is in picking sentences with a certain linguistic meaning that constrains the range of available literal interpretations. What questions the hearer wants to answer by engaging in discourse and what background beliefs she has determine whether she will include a PoR feature in her interpretative mental model of an attitude report sentence. The question "Did the speaker intend to convey this way of representing so-and-so?" arises only if the hearer is especially concerned with what exactly is on the speaker's mind. Nothing I have said rules out that we often *do* particularly care about recovering the proposition that the speaker specifically intended to bring across. The claim is simply that reflections about a speaker's intentions are not *by default* driving literal utterance interpretation.

General utterance interpretation as the recognition of particular communicative intentions is not "cognitively encapsulated" in Fodor's sense, so a strong predictive theory is highly unlikely in this area.[61] Yet, as Atlas has argued, we *can* give a theory of default literal interpretations with some predictive success, if such a theory is not directly dependent on guessing intentions.[62] I have presented such a theory with the Open Instruction Theory combined with (IMRI).

## 5. Contrasting Discourse Representation Theory and "hidden anaphora"

The framework of Discourse Representation Theory (DRT) shares with the Open Instruction Theory the view that in order to make sense of utterance interpretation, we need a level of representation beyond what is generated by syntax and lexical items. Now, Asher and Kamp's DRT account of attitude report sentences could be characterized as a "hidden anaphora" theory.[63] The starting point for this theory is the notion that the use of a proper name presupposes that a reference marker standing for the referent of the name already exists in the representation of the discourse, where this antecedent reference marker must be linked to a set of conditions for determining the referent. Now, if a proper name occurs within the that-clause of a belief report, the use of the proper name presupposes that the hearer has an antecedent reference maker standing for the referent of the proper name *within a dedicated representation of the believer's total cognitive state*.[64] Asher holds that, "from a DR theoretic perspective, the goal of the speaker is to get the recipient to approximate the true picture (insofar as it is known to the speaker) of the target Belief and its internal connections to other components of the believer's cognitive state".[65] If there is no antecedent representation representing reference markers in the believer's cognitive state, we have to *accommodate* the existence of such a representation, including, on Asher's view, connections to "schematic"

59. Blackburn (1994), p. 62

60. It is a separate question what principles govern conversational implicatures beyond what is involved in literal interpretation and how those principles relate to perceived intention.

61. Fodor (1983)

62. Atlas (2005)

63. Asher (1986; 1993); Kamp (1988; 1990)

64. Asher (1986), p. 144. Also see Kamp (1990), p. 41–87. Asher's position seems to be that a proper name in an attitude report context should introduce two presuppositions: one for the hearer and one for the target of the belief report. I am not sure whether Kamp would say that the belief report "shifts" the presupposition of the proper name to a presupposition about the believer or that the belief report generates an additional presupposition.

65. Ibid.

representations of the believer's conditions for determining the referent.[66] Different ways of representing an individual would roughly correspond to different internal connections to other components of the believer's cognitive state. Once we have at least *accommodated* a representational token in a representation of someone's total cognitive state, we get an interpretation of a proper name in an attitude report by resolving the anaphora postulated in the proper name to that token.

This view faces a version of the meaning-intention problem discussed in the introduction. While there is a sense in which the use of proper names tends to presuppose familiarity with the referent,[67] it just does not seem right to say that proper names in the that-clauses of attitude reports carry the presupposition of an antecedent representation of the cognitive state of the bearer of a reported attitude. Suppose Mike sees an anonymous valentine in John's mailbox and tells his neighbor,

(42) Someone likes John.[68]

It is implausible that this utterance in any way presupposes that John's neighbor has an antecedent representation of "someone's" mental state, any more than Mike's utterance conveys a certain type of way of representing John. On the DRT account, there are no default singular interpretations, because the core idea is that we have to tie our interpretations to an existing representation of the target's total cognitive state. Instead of default singular interpretations, you have a form of accommodation if the presupposition of an existing representation of the target's total cognitive state fails. However, what I argued in the introduction, in seeming agreement with Schiffer and Jascszolt, was that unless there are special contextual features that prompt us to do otherwise, our interpretations of attitude reports are concerned with the environment rather than directly with the peculiarities of people's cognitive states. As I argued with respect to various examples, even if we *do* have additional information about peculiarities of people's cognitive states, we still go for singular interpretations by default. If we had the presupposition that Asher and Kamp postulate, it would be surprising if we got default singular interpretations even when we *are* in a position to take into account the internal connections of the beliefs of the target of the attitude report. To involve a representation of anything close to someone's "total cognitive state" in interpreting an attitude report seems to be the exception rather than the norm, so it seems wrong to build a presupposition of such a representation into attitude reports.

Even if we grant that, with a suitably revised version of Asher and Kamp's view that relies less on conditions for determining referents instead of PoRs, a correctly accommodated presupposition can fully mimic a default singular interpretation in terms of the resulting content of the utterance in context, there is still a difference between a theory that gets the result via *accommodation* and a theory that gets it via *default interpretation*. Since psycholinguistics has begun to shed some light on the processing costs of presupposition failure and accommodation, the relevant difference between my view and the DRT theory of attitude reports is amenable to experimental inquiry. Violations of presuppositions followed by accommodation yield longer reading times on the critical word that engenders the presupposition, as empirically demonstrated by Tiemann et al. (2011). My view predicts the same default interpretations in various cases independently of the availability of background information about the cognitive states of the target. Asher and Kamp's view predicts presupposition failure and subsequent accommodation for cases without relevant background information, but not for cases with the right background information. Assuming Tiemann et al.'s results, this means that Asher and Kamp's view predicts a difference in reading times between certain cases that one could construct, where my theory predicts no such difference. Designing a carefully matched set of cases for an actual experiment

66. Asher (1986), p. 145; Kamp (1987), p. 172

67. See Heim (1982).

68. I take for granted that the relevant sense of 'likes' behaves like an attitude relation. Cf. 'Lois likes Superman, but she only tolerates Clark Kent.'

would be a difficult task, but the point is that the debate is amenable to experimental inquiry in principle.

A general worry about DRT approaches, raised by Levinson (2000), is that there is a lack of predictive power about which interpretations will be dominant in various types of contexts.[69] Different interpretations of attitude report sentences are treated as different ways of resolving anaphora (hence the commitment to a presupposition of a suitable range of antecedents), without a fully systematic decision procedure between competing interpretations. By contrast, one of the key motivations of the theory I have presented was to account for what seemed to be striking regularities in how context influences the interpretation of attitude report sentences and in the sorts of intuitions generated by "null" or default contexts. Moreover, every effort was made to build those parts of OIT that are not syntactically constrained out of components that have been independently postulated in the psychology of reasoning, while little effort seems to have been made to concretely motivate the complex machinery of DRT independently. That said, if the worries just described are properly taken into account, there appears to be no in-principle barrier to integrating various insights DRT has yielded into various other fragments of language with the Open Instruction Theory of attitude reports.

## 6. Interpretation shifts across VP ellipsis

As noted in the introduction, there are different ways in which the linguistic meaning of attitude report sentences could relate to the various types of interpretations that we can obtain in different contexts. On the Open Instruction Theory, as on the views proposed in recent work by Soames and Bach, the linguistic meaning is silent on the difference between these interpretations; it is nonspecific with respect to them.

Nonspecificity theories contrast with views on which attitude report sentences include a constituent that is *indexical* with respect to types of ways of representing, just as, say, 'that' is indexical with respect to an individual referent. An interesting feature of indexical expressions is that one cannot shift their interpretation across various ellipsis constructions. Consider the following sentences with uncontroversial indexical expressions:

(43) Lois owns that, and so does Mary.

(44) [Lois hates only Scorsese, and Mary hates only Lukas] Lois hit him, and so did Mary.

(45) Lois came yesterday, but Mary did not.

The indexical in the elided VP cannot independently contribute a novel referent in interpretation. We have to interpret (43) as claiming that Lois and Mary own the same thing (or kind of thing), (44) as claiming that they hit the same person, and (45) as claiming that Mary came on a different day. That these constraints exist is not particularly controversial, though there is an open question as to what general principle accounts for them. At least part of the constraint seems to be due to a grammatical requirement that the elided VP has to be syntactically "parallel" to its antecedent VP in certain ways.[70] On a classical view on which syntax encodes indices for indexicals (or some functional equivalent), a version of this principle can explain the constraints on interpretation in the examples. In contrast to indexicals, certain other kinds of expressions that also involve interpretations that can vary in context do not seem to incur this constraint on interpretation under ellipsis.

A non-indexical kind of context sensitivity is plausibly involved in the following example:

(46) [Mike is throwing a party in New York; Bill is throwing a party in Boston.]

69. Levinson (2000), p. 248

70. Fiengo and May (1994)

(47) Mike hopes Mia will come, and so does Bill. One of them will be disappointed.[71]

Here, Mike and Bill are characterized as hoping that Mia will arrive at different locations, even though 'hopes Mia will come' is elided. While it is possible to shift relevant aspects of context across ellipsis constructions, indexical expressions do not seem to be able to shift their contribution across ellipsis, in contrast to cases like (47). There are grammatical constraints on how indexicals can be interpreted under ellipsis, regardless of whether we have a shift of context mid-sentence.

A theory like the Open Instruction Theory that characterizes attitude report sentences as nonspecific with respect to types of ways of representing predicts that it should be possible to find cases in which we switch our interpretation of a VP like 'believes that $F(x)$' with respect to types of ways of representing across ellipsis, given a suitable context. In other words, we expect that, with respect to types of ways of representing, attitude report sentences should pattern with 'come', rather than with 'that' or 'him'.

Consider the following scenario: Spielberg and Coppola have an unshakable belief that Hathaway is a talented Shakespearean actress, and both invite her to star in their respective upcoming rival productions of Shakespeare's *Twelfth Night*. As a personal challenge, Hathaway decides to audition incognito for both Spielberg and Coppola, appearing as Viola for Spielberg and as Sebastian for Coppola. We know about all of this and have an informant who spies on both auditions. Hathaway has a bad day and does well for Spielberg but not for Coppola. Our informant reports,

(48) Spielberg believes that Hathaway is talented, but Coppola does not <believe that Hathaway is talented>.

In the context provided, (48) seems true and intuitively conveys that Coppola and Spielberg represent Hathaway differently: Spielberg believes that Hathaway, *presented as Viola*, is talented, while Coppola does not believe that Hathaway, *presented as Sebastian*, is talented. Of course, they both believe of Hathaway that she is talented.

Consider another scenario: Like most ordinary Gotham City criminals, mafia boss Zucco is terribly afraid to meet Batman. However, he is very keen to meet Bruce Wayne, because he wants to convince him to join his Ponzi scheme. He does not know Bruce Wayne is Batman. By contrast, The Riddler has no interest in Wayne but wants to challenge Batman to a fight. Bruce's butler, who knows all of this, sums it up:

(49) The Riddler hopes to meet Master Wayne, and so does Zucco.

Intuitively, (49) can convey in context that The Riddler bears an attitude toward Wayne represented in one way, while Zucco bears an attitude toward Wayne represented in another way.

Examples like (48) and (49) suggest that the relevant differences in how we can interpret 'believes that $F(x)$' in context is more similar to the sort of context sensitivity we find in 'come' than to more grammatically constrained context sensitivity of indexicals. The availability of interpretations like the one we can obtain from (48) in context is, at the very least, *surprising* on the hypothesis that attitude report sentences are indexical with respect to types of ways of representing. On the Open Instruction Theory, the linguistic meaning is silent on types of ways of representing, so it is not surprising that we do not find relevant grammatical constraints on interpretation. Note that the issue is not whether contexts can shift mid-sentence; they clearly *can* be shifted. The issue is that some changes in interpretation seem to be grammatically blocked regardless of context while others are not. Attitude report

71. If one is worried about the possibility of a reflexive interpretation of 'come', one could change the scenario to: 'Mike is throwing a party in New York; Bill is in Timbuktu, but his son is celebrating his 21st birthday in Boston. Mike hopes Mia will come, and so does Bill. One of them will be disappointed.' Contrast this with 'Mike is throwing a party in New York; Bill is in Timbuktu, but his son is celebrating his 21st birthday in Boston. #Mike hopes Mia will be his guest, and so does Bill [understood in a crossed way].'

sentences do not seem to relevantly pattern with indexical expressions in this regard.

Schiffer has explored other syntactic arguments against indexicalism about attitude reports, prompting various revisions to his view.[72] I argue elsewhere that the intuitions discussed above present a challenge to various versions of indexicalism that seems surprisingly hard to avoid by technical modifications, but I do not have the space to pursue this issue here.[73] For the purposes of this paper, I simply want to observe that it is a virtue of the Open Instruction Theory that it makes it unsurprising that the way attitude reports behave under VP ellipsis patterns with cases like 'come' rather than with cases like 'that', 'him', etc.

## 7. The overgeneration objection

One general objection that has been raised against theories that do not derive all differences in interpretation from associated differences in semantics is that those theories overgenerate interpretations. For example, Stanley has attacked so-called "free pragmatic enrichment" theories on these grounds.[74] Unlike the "free enrichment" tradition, OIT directly starts with the problem of giving a systematic account of linguistic meaning. The linguistic meaning of a sentence is fully compositionally determined by rules continuous with generative grammar, and the interpretations at issue are still *literal interpretations*. The range of literal interpretations is systematically determined by linguistic meaning, not just assumed; there is no reliance on a general-purpose enrichment process. The pragmatic component of the theory makes clear predictions about how particular features of context drive particular interpretations and appears to predict correctly the default interpretations of attitude report sentences as well as the range of non-default interpretations.

## 8. Mental models, content, and truth conditions

On the Open Instruction Theory, attitude report sentences, taken by themselves, do not have truth conditions. Their contributions are partially open instructions to construct mental models, which constitute possible interpretations of utterances of attitude report sentences. Mental models express propositions that have truth conditions. We may identify the truth conditions of the model with the truth conditions of the proposition it expresses. When an utterance of an attitude report sentence can be associated with a correct interpretation—that is, a correct kind of mental model—(or if all correct models have the same truth conditions), we identify the truth conditions of that *utterance* with the truth conditions of its correct mental model(s).

Not every sentence in natural language can be associated with unique truth conditions. On this point, the Open Instruction Theory is in agreement with early theorists like Atlas and Chomsky who held that linguistic meanings are often semantically nonspecific, as well as with theorists like Bach and Soames.[75] At the same time, we give specific interpretations to utterances of semantically non-specific sentences, because (IMRI) assigns interpretive mental models with specific truth conditions. By default, the interpretation (IMRI) assigns

72. Schiffer (1992); Ludlow (1995)

73. See Koralus (2010). I will just briefly note that it will not do, for example, to postulate a reflexive constituent rather than an indexical, on the analogy of 'his mother'. It is clearly possible to interpret 'John likes his mother, and so does Jack' as saying that John likes John's mother while Jack likes Jack's mother. One then might try to circumvent the ellipsis problem by abandoning indexicalism in favor of "reflexical-ism" about attitude report sentences, which would roughly involve paraphrases like 'Peter1 believes that Paderewski can fly, involving his1 PoR.' Figuring out the referent of 'his PoR' is rather different from fig-uring out the referent of, e.g., 'John's mother'. John presumably has only one mother. By contrast, Peter has at least two types of ways of representing Paderewski for the individual Paderewski, and he could have arbitrarily many. Which PoR enters into an interpretation of an attitude report sentence depends on context. Given the dialectical situation, it will not do to say that 'his PoR' is ambiguous or indexical as well as re-flexive. This would bring us back to the VP ellipsis problem. A fuller discussion can be found in Koralus (2010), where I also consider and reject the proposal that the interpretation at issue could be captured if the indexical resides on what syntacticians call "little" vP.

74. Stanley (2005); Hall (2008); Carston and Powell (2006)

75. Atlas (1977; 2005); Chomsky (2000); Soames (2004); Bach (2000)

is the best interpretation relative to our background beliefs and the question we are trying to answer by engaging in discourse, which may not be what we have best reason to think is *intended* by the speaker. By default, our background questions tend to be directed toward the environment and toward what people take their environment to be like, rather than directed toward the mental states by means of which they represent the environment. However, if it is part of our background question that we want to know specifically what the speaker thinks, then (IMRI) allows for the possibility of different results. Moreover, nothing I have said rules out that once we have a literal utterance interpretation, traditional Gricean pragmatics may be responsible for generating conversational implicatures.

## 9. The unity of cognitive science: from the principle of linguistic constraint to the principle of psychological constraint

One of the main attractions of accounts of literal utterance interpretation purely in terms of the semantics of lexical items and their mode of combination is that their interaction with syntactic constraints can often yield surprising predictions. Purely syntactic observations are an independent source of data that constrain such accounts. As I have argued, along with Atlas, Bach, Chomsky, and others, not all aspects of interpretation, even "literal interpretation", are going to yield to an account purely in these terms.

I suggest that the lesson to be drawn is not that legitimate theorizing is possible only if we stipulate that every difference in interpretation has to be traceable to a difference in syntax and lexical items. Rather, the lesson should be that we should proceed in a way that puts us in contact with independent sources of data. This paper showed how cognitive psychology can be used to independently motivate principles of a theory of utterance interpretation beyond syntax and lexical items. If we are discussing differences in interpretation that are not reflected in syntax, we should find different sources of evidence for the mechanisms we invoke. Cognitive psychology is a good candidate source for independent evidence. The most promising path for the theory of utterance interpretation beyond syntax and the lexicon is a path toward the unity of cognitive science.

## References


Asher, N. (1986). "Belief in discourse representation theory". *Journal of Philosophical Logic* 15(2), pp. 127–189.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer.

Asher, N., and Lascarides, A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Atlas, J. (1977). "Negation, ambiguity, and presupposition". *Linguistics and Philosophy* 1(3), pp. 321–336.

Atlas, J. (2005). *Logic, Meaning, and Conversation: Semantical Underdeterminacy, Implicature, and Their Interface*. Oxford: Oxford University Press.

Atlas, J., and Levinson, S. (1981). "*It*-Clefts, Informativeness, and Logical Form: Radical Pragmatics (Revised Standard Version)". In: Peter Cole (ed.), *Radical Pragmatics*. New York: Academic Press.

Bach, K. (2000). "Do Belief Reports Report Beliefs?" In: K. M. Jaszczolt (ed.), *The Pragmatics of Propositional Attitude Reports*. Oxford: Oxford University Press.

Barrouillet, P., Grosset, N., and Lecas, J. (2000). "Conditional reasoning by mental models: chronometric and developmental evidence". *Cognition*, 75(3), pp. 237–266.

Bauer, M. I. and Johnson-Laird, P. N. (1993). "How Diagrams Can Improve Reasoning". *Psychological Science* 4(6), pp. 372–378.

Blackburn, S. (1994). *The Oxford Dictionary of Philosophy*. Oxford: Oxford University Press.

Braun, D., and Saul, J. (2002). "Simple Sentences, Substitutions, and Mistaken Evaluations". *Philosophical Studies* 111(1), pp. 1–41.

Byrne, R. (2005). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge: MIT Press.

Carston, R., and Powell, G. (2006). "Relevance Theory—New Directions and Developments". In: Lepore, E., and Smith, B. (eds.), *The Oxford Handbook of Philosophy of Language*. Oxford: Oxford University Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge: MIT Press.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Crimmins, M. (1992). "Context in the attitudes". *Linguistics and Philosophy* 15(2), pp. 185–198.

Fiengo, R., and May, R. (1994). *Indices and Identity*. Cambridge: MIT Press.

Fodor, J. (1983). *The Modularity of Mind*. Cambridge: MIT Press.

Groenendijk, J., and Stokhof, M. (1991). "Dynamic Predicate Logic". *Linguistics and Philosophy* 14 (1), pp. 39–100.

Hall, A. (2008). "Free enrichment or hidden indexicals?". *Mind and Language* 23(4), pp. 426–456.

Hamblin, C. (1973). "Questions in Montague English". *Foundations of Language* 10(1), pp. 41–53.

Harman, G. (1965). "The Inference to the Best Explanation". *The Philosophical Review* 74(1), pp. 88–95.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts Amherst.

Heim, I., and Kratzer, A. (1998). *Semantics in Generative Grammar*. Malden, MA: Blackwell.

Hintikka, J. (1969). "Semantics for propositional attitudes". In: J. Davis et al. (eds.), *Philosophical Logic*. Dordrecht: D. Reidel.

Hobbs, J. R., Stickel, M., Appelt, D., and Martin, P. (1993). "Interpretation as abduction". *Artificial Intelligence* 63(1–2), pp. 69–142.

Jaszczolt, K. M. (2000). "The default-based context-dependence of belief reports". In: K. M. Jaszczolt (ed.), *The Pragmatics of Propositional Attitude Reports*. New York: Elsevier.



Johnson-Laird, P. N. (1983). *Mental models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (2008) "Mental models and deductive reasoning". In: L. Rips and J. Adler (eds.), *Reasoning: Studies of Human Inference and Its Foundations*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., and Caverni, J. (1999). "Naïve Probability: A Mental Model Theory of Extensional Reasoning". *Psychological Review* 106(1), pp. 62–88.

Kamp, H. (1981). "A theory of truth and semantic representation". In: J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof (eds.), *Formal Methods in the Study of Language*. Amsterdam: Mathematical Centre Tracts 135, pp. 277–322.

Kamp, H. (1987). "Comments on Stalnaker, Belief Attribution and Context". In: R. H. Grimm and D. D. Merrill (eds.), *Contents of Thought*. Tucson: University of Arizona Press.

Kamp, H. (1990). "Prolegomena to a structural theory of belief and other attitudes". In: C. A. Andersonand J. Owens (eds.), *Propositional Attitudes: The Role of Content in Language, Logic, and Mind*. Stanford: CSLI Publications.

Kaplan, D. (1989). "Demonstratives". In: J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan*. Oxford: Oxford University Press.

Knauff, M., Mulack, T., Kassubek, J., Salih, H. R., and Greenlee, M. W. (2002). "Spatial imagery in deductive reasoning: a functional MRI study". *Cognitive Brain Research* 13(2), pp. 203–212.

Koralus, P. (2010). *Semantics in Philosophy and Cognitive Neuroscience: The Open Instruction Theory of Attitude Report Sentences, Descriptions, and the Necker Cube*. Doctoral dissertation, Princeton University.

Koralus, P. (2011). "Descriptions, Ambiguity, and Representationalist Theories of Interpretation". *Philosophical Studies* [Epub ahead of print doi: 10.1007/s11098–011–9759–5].

Koralus, P. (*in preparation*). "Axiomatic Mental Model Theory".

Kripke, S. (1979). "A Puzzle about Belief". In: A. Margalit (ed.), *Meaning and Use*. Dordrecht: D. Reidel.

Kroger, J. K., Nystrom, L. E., Cohen, J. D., and Johnson-Laird, P. N. (2008). "Distinct neural substrates for deductive and mathematical processing". *Brain Research* 1243, pp. 86–103.

Lansky, B. (1999). *Baby Names Around the World*. Minnetonka: Meadowbrook Press.

Lewis, D. (1979). "Scorekeeping in a Language Game". *Journal of Philosophical Logic* 8(1), pp. 339–359.

Levinson, S. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge: MIT Press.

Ludlow, P. (1995). "Logical form and the hidden-indexical theory: a reply to Schiffer". *The Journal of Philosophy* 92(2), pp. 102–107.

Mascarenhas, S. (2009). *Inquisitive Semantics and Logic*. M.Sc. thesis. Amsterdam: Institute for Logic, Language and Computation.

McGlone, M. (2007). Doctoral dissertation, Princeton University.

Muskens, R. (1996). "Combining Montague Semantics and Discourse Representation". *Linguistics and Philosophy* (19), pp. 143–186.

Nichols, S., and Stich, S. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds.* Oxford: Clarendon Press.

Pylkkänen, L., and McElree, B. (2007). "An MEG study of silent meaning". *Journal of Cognitive Neuroscience* 19(11), pp. 1905–1921.

Récanati, F. (1993). *Direct Reference: From Language to Thought*. Oxford: Blackwell.

Récanati, F. (2002). "Unarticulated constituents". *Linguistics and Philosophy* 25(3), pp. 299–345.

Richard, M. (1990). *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them*. Cambridge: Cambridge University Press.

Rips, L. J. (1983). "Cognitive Processes in Propositional Reasoning". *Psychological Review* 90(1), pp. 38–71.

Schiffer, S. (1977). "Naming and Knowing". *Midwest Studies in Philosophy* 2(1), pp. 28–41.

Schiffer, S. (1992). "Belief Ascription". *The Journal of Philosophy* 89(10), pp. 499–521.



Schiffer, S. (2000). "Propositional attitudes in direct-reference semantics". In: K. M. Jaszczolt (ed.), *The Pragmatics of Propositional Attitude Reports*. Oxford: Oxford University Press.

Soames, S. (1987). "Substitutivity". In: J. J. Thomson (ed.), *On Being and Saying: Essays for Richard Cartwright*. Cambridge, MA: MIT Press.

Soames. S. (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity.* Oxford: Oxford University Press.

Soames, S. (2004). "Naming and Asserting." In: Zoltan Szabo (ed.), *Semantics versus Pragmatics*. Oxford: Oxford University Press.

Stanley, J. (2002). "Making it articulated". *Mind & Language* 17(1–2), pp. 149–168.

Stanley, J. (2005). "Semantics in Context". In: G. Preyer and G. Peter (eds.), *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford: Oxford University Press.

Tiemann, S., Schmid, M., Bade, N., Rolke, B., Hertrich, I., Ackermann, H., Knapp, J., and Beck, S. (2011). "Psycholinguistic Evidence for Presuppositions: On-line and Off-line Data". In: I. Reich et al. (eds.), *Proceedings of Sinn & Bedeutung* 15. Saarbrücken: Saarland University Press.

Zwicky, A. M., and Sadock, J. M. (1975). "Ambiguity tests and how to fail them". In: J. P. Kimball (ed.), *Syntax and Semantics* 4. New York: Academic Press.