Hyperfine-grained Meanings in Classical Logic*

Reinhard Muskens

1 Logical Omniscience

Let us call two expressions *synonymous* if and only if they may be interchanged in each sentence without altering the truth value of that sentence.[1] With the help of an argument by Benson Mates (Mates [1950]) it can be shown that synonymy is a very strong relation indeed. Consider, for example, the following two sentences.

(1) Everybody believes that whoever thinks that all Greeks are courageous thinks that all Greeks are courageous

(2) Everybody believes that whoever thinks that all Greeks are courageous thinks that all Hellenes are courageous

Some philosophers indeed believe that whoever thinks that all Greeks are courageous also thinks that all Hellenes are courageous.[2] But certainly not everyone agrees, and so (2) is false. We may assume, on the other hand, that (1) is true, and since (1) can be obtained from (2) by replacing 'Hellenes' by

*This is a reprint from *Logique & Analyse* **133-134**, 1991, pp. 159-176. A preliminary version of the paper appeared as 'Logical Omniscience and Classical Logic' in D. Pearce and G. Wagner (eds.), *Logics in AI,* Lecture Notes in Artificial Intelligence **633**, Springer, Berlin, 1992, 52-64. I would like to thank Ed Keenan, Heinrich Wansing and the audiences at JELIA '92 and the Epistemic Logic Colloquium for comments and criticisms.

[1] This essentially is Mates's [1950] formulation of the interchangeability principle. Note how close Mates's formulation is to Leibniz's:

Eadem sunt quorum unum potest substitui alteri salva veritate.

[2] E.g. Putnam in Putnam [1954].

'Greeks', the latter two words, somewhat surprisingly, are not synonymous. By a similar procedure any pair of words that are normally declared synonyms can be shown not to be synonymous after all, if our definition of the term is accepted.

Worse, it seems that the relation of synonymy is even stronger than the relation of logical equivalence is. Sentences that are normally accepted to be logically equivalent need not be synonymous. Suppose[3] Jones wants to enter a building that has three doors, A, B, and C. The distances between any two of these doors are equal. Jones wants to get in as quickly as possible, without making detours and he knows that if A is locked B is not. Now, if our agent tries to open door B first and finds it locked there might be a moment of hesitation. The reasonable thing for Jones is to walk to A, since if B is locked A is not, but he may need some time to infer this. This contrasts with the case in which he tries A first, since if he cannot open this door he will walk to B without further ado. The point is that one may well fail to realize (momentarily) that a sentence is true, even when one knows the contrapositive to hold. For a moment (3) might be true while (4) is false.

(3) Jones knows that if A is locked B is not locked

(4) Jones knows that if B is locked A is not locked

It follows that the two embedded sentences are not synonymous, even though logically equivalent on the usual account.

All reasoning takes time. This means that

(5) Jones knows that φ

need not imply

(6) Jones knows that ψ,

even if φ and ψ are logically equivalent. If the embedded sentences are syntactically distinct then, since Jones needs time to make the relevant inference, there will always be a moment at which (5) is true but (6) is still false.

Of course, the easier it is to deduce ψ from φ the shorter this time span will be and the harder it is to imagine that Jones knows φ without realizing

[3] This example is adapted from Moore [1989].

that ψ. We ascribe to Jones certain capacities for reasoning, even if we do not grant him logical omniscience. For instance, if Jones knows A and B, then it is hard to imagine that he would fail to know B and A as well. But even here there may be a split-second where the necessary calculation has yet to be made. Therefore 'Jones knows B and A' does not follow from 'Jones knows A and B'.

A real problem results if we try to formalise the logic of the verb *to know* and its like. All ordinary logics (including modal logics) allow logical equivalents to be interchanged, but epistemic contexts do not admit such replacements. It thus may seem that the logic of the propositional attitude verbs is very much out of the ordinary and that we must find a logic that does not support full interchangeability of equivalents if we want our theory of the propositional attitudes to fit the facts.

Systems that do not admit replacement of equivalents do in fact exist. For example Rantala [1982^a, 1982^b], working out ideas of Montague [1970], Cresswell [1972] and Hintikka [1975], offers an 'impossible world semantics' for modal logic in which the interchangeability property fails. In the next section I'll criticise Rantala's system for not being a *logic* in the strict sense, but I think that its main underlying idea, the idea that we can use 'impossible' worlds to obtain a very fine grained notion of meaning, is important and useful. Despite appearances however, this idea is compatible with classical logic and in section 4 I shall show in some detail how we can use impossible worlds to treat the propositional attitudes without resorting to a non-standard logic. The logic that I shall use is the classical type theory of Church [1940]. Section 3 will be devoted to a short exposition of this logic for the convenience of those readers who are not already familiar with it.

2 Rantala Models for Modal Logic

The basic idea behind Rantala's interpretation of the language of propositional modal logic is to add a set of so-called impossible (or: non-normal) worlds to the usual Kripke frames. Equivalence can then be defined with respect to possible worlds only, but for interchangeability the impossible ones come into play as well. The net result will be that equivalents need not be interchangeable in the scope of the epistemic operator K.

Formally, a *Rantala model* for the language of propositional modal logic is a tuple $\langle W, W^*, R, V \rangle$ consisting of a non-empty set W of possible worlds, a set W^* of impossible worlds, a relation $R \subseteq (W \cup W^*)^2$, and a two-place valuation function $V : \textsc{Form} \times (W \cup W^*) \to \{0,1\}$ such that for all $w \in W$:

(i) $V(\neg\varphi, w) = 1$ iff $V(\varphi, w) = 0$

(ii) $V(\varphi \wedge \psi, w) = 1$ iff $V(\varphi, w) = 1$ and $V(\psi, w) = 1$

(iii) $V(K\varphi, w) = 1$ iff $V(\varphi, w') = 1$ for all $w' \in W \cup W^*$ such that wRw'.

Note that the value of a complex formula in an impossible world can be completely arbitrary. The value of a complex formula does not depend on the values of its parts. Note also that Rantala models clearly generalize Kripke models: a Kripke model simply is a Rantala model with empty W^*.

A formula ψ is said to *follow from* a formula φ if $V(\varphi, w) = 1$ implies $V(\psi, w) = 1$ in each Rantala model $\langle W, W^*, R, V\rangle$ and $w \in W$. A formula φ is *valid* in a Rantala model if $V(\varphi, w) = 1$ in each $w \in W$. A formula φ is *valid* simpliciter if it is valid in each Rantala model. Formulae φ and ψ are *equivalent* iff φ follows from ψ and ψ follows from φ. Clearly, all propositional tautologies are valid, but Necessitation fails: validity of φ does not imply validity of $K\varphi$. This is as it should be: one may well fail to know that a sentence is true even if it is valid. Also the K schema fails: write $\varphi \to \psi$ for $\neg(\varphi \wedge \neg\psi)$, then $K(\varphi \to \psi) \to (K\varphi \to K\psi)$ is not valid. This is also as desired, since knowledge is not closed under modus ponens. The notion of validity just defined is indeed a minimal one: with the help of standard techniques it is easily shown that a sentence is valid if and only if it is a substitution instance of a propositional tautology.

Equivalent sentences need not be interchangeable in this system. For instance, p and $\neg\neg p$ are equivalent, but a model in which Kp and $K\neg\neg p$ are assigned different truth values in some possible world is easily constructed. The system thus meets the requirement discussed in the previous section.

Wansing [1990] shows that a number of formalisms that have arisen in AI research can in fact be subsumed under the impossible worlds approach. He proves in some detail that the knowledge and belief structures that were proposed in Levesque [1984], Vardi [1986], Fagin & Halpern [1988] and Van der Hoek & Meyer [1988] can be reduced to non-normal worlds models. Thijsse [1992] on the other hand proves that Rantala models can be reduced to the models of Fagin & Halpern [1988] (if the latter are slightly generalised) and thus, using Wansing's result, obtains equivalence between Rantala semantics and Fagin & Halpern's logic of general awareness.

The system is elegant enough, generalises Kripke's semantics for modal logic and subsumes other approaches to the logic of the propositional attitudes, then what are my qualms? Here is one. We have just defined the implication $\varphi \to \psi$ with the help of $\neg$ and $\wedge$. Alternatively, we could have introduced the arrow as a primitive, imposing the extra constraint that for all models $\langle W, W*, R, V\rangle$ and for all $w \in W$:

(ii′) $V(\varphi \Rightarrow \psi, w) = 0$ iff $V(\varphi, w) = 1$ and $V(\psi, w) = 0$.

But the two methods lead to different results. For even while the formulae $\varphi \to \psi$ and $\varphi \Rightarrow \psi$ are equivalent, they are not interchangeable in all contexts: $K(\varphi \to \psi)$ and $K(\varphi \Rightarrow \psi)$ may have different truth values in some possible world. The addition of $\Rightarrow$ to the language and the addition of clause (ii′) to the definition of a Rantala model really added to the logic's expressive power. Two Rantala models may validate exactly the same sentences of the original language, yet may differ on sentences of the new language.

This means that functional completeness fails for Rantala's system. Usually, when setting up a logic, we can contend ourselves with laying down truth conditions for some functionally complete set of connectives, say $\neg$ and $\wedge$. Adding connectives and letting them correspond to new truth functions usually does not increase expressive power since all truth functions are expressible with the help of $\neg$ and $\wedge$. But in the present case this is no longer so.

Why? The source of the trouble is that in Rantala models the interpretation of logical constants is not fixed. Even a reduplication of one of the logical constants would strengthen the system. Let us add a connective & to the system and impose a condition completely analogous to (ii), namely that for all $w \in W$:

(ii″) $V(\varphi \mathbin{\&} \psi, w) = 1$ iff $V(\varphi, w) = 1$ and $V(\psi, w) = 1$

The weird result is that $\mathrm{K}(\varphi \wedge \psi)$ is *not* equivalent with $\mathrm{K}(\varphi \mathbin{\&} \psi)$ and that it is possible now to distinguish between models that were indistinguishable (i.e. validated the same sentences) before. The question arises: which is the real conjunction, $\wedge$ or &?

Is it possible to have a *logic* if the interpretations of the logical constants are allowed to vary with each model? This question can only be answered if some criterion of logicality is accepted. Such criteria have been developed within abstract model theory (see Barwise [1974]), a branch of logic where theorems are proved of the form: "every logic that has such-and-such properties is so-and-so".[4] Rantala [1982^b] notes that in fact his system does not meet the standards that are usually set here. There is a problem with *renaming.* In general the truth value of a formula should not change if we replace some non-logical constant in it by another constant which has the same semantic value. In Rantala models this fails, for example, it is easy to construct a Rantala model such that $V(p, w) = V(q, w)$ for all $w \in W \cup W^*$ but $V(K\neg p, w) \neq V(K\neg q, w)$ for some possible world $w \in W$. The value of $K\neg p$ thus may crucially depend on the particular name that we have chosen for the proposition that is denoted by p.

This may or may not be defensible, but, as I shall show below, the weird characteristics of Rantala's system are not essential to the idea of impossible world semantics. The idea can be formalised with the help of a system that meets all standards of logicality.

The basic intuition behind the introduction of impossible worlds is that, since we humans are finite and fallible, we fail to rule out worlds which would be ruled out by a perfect reasoner. What do such worlds look like? Well, for example, one of Jones' epistemic alternatives was the impossibility that 'if A is locked B is not' is true, but that 'if B is locked A is not' is false. For a short time, at least one impossible world in which the first sentence is true but the second is not was not ruled out by Jones' reasoning. But in such worlds the words 'not' and 'if' cannot get their usual Boolean interpretation, since this interpretation would simply force the sentences to be equivalent. We therefore end up with non-standard interpretations for the 'logical' words in English: 'and' cannot be intersection of sets of worlds, 'or' cannot be union, 'not' cannot be complementation and so on.

Rantala formalises this by treating the word 'and' as the connective $\wedge$ but by giving this last symbol a non-logical interpretation. This leads to a funny system. The obvious alternative is to keep the logic standard but to formalise the English word 'and' and its like as non-logical constants: once it is accepted that no *logical* operation strictly corresponds to the English word 'and', the most straightforward solution is be open about it and to formalise the word with the help of a non-logical constant.

Of course, some connection between the 'logical' words in English and

[4] An example is Lindström's Theorem, which says that for no logic properly extending first-order predicate logic both the Compactness Theorem and the Löwenheim-Skolem Theorem hold.

the connectives that usually formalise them should remain intact. What connection? Even if we allow the interpretations of the 'logical' words to be completely arbitrary, there will be a subset of the set of all worlds where 'and' and its ilk behave standardly. These worlds where the logical words of English have their usual logical interpretation may be called the 'possible' or 'actualizable' ones. As we shall see below, the assumption that the actual world is actualizable leads to the desired relation of logical consequence.

But it is high time for a more precise formalisation. In the next section I give a short sketch of the classical logic that I want to use and in the last section I'll apply it to the propositional attitude verbs.

3 Classical Type Theory

Since we want to treat 'and', 'or', 'not', 'if', 'every' and 'some' as non-logical constants, we should use a logic that admits of non-logical constants for these types of expressions. Ordinary predicate logic will not do, but a logic that is admirably suited to the job is Church's [1940] formulation of Russell's Theory of Types (Russell [1908]). Since I expect that not all of my readers are familiar with this system, I'll give it a short exposition (and so readers who already know about the logic can skip this section). For a more extensive account one may consult the original papers (e.g. Church [1940], Henkin [1950, 1963]), Gallin [1975], the survey article Van Benthem & Doets [1983], or the text book Andrews [1986]. In Muskens [1989^a, 1989^b, 1989^c] some variants of the logic are given, but I'll follow the standard set-up here.

In classical type theory each logical expression comes with a *type*. Types are either basic or complex. The type of truth values, here denoted with t, should be among the basic types, but there may be other basic types as well. In this paper, for example, we'll assume types for individuals (type e) and worlds (type s). Complex types are of the form $\alpha\beta$[5] and an expression of type $\alpha\beta$ will denote a function which takes things of type α to things of type β. Formally we define:

Definition 1 (Types). The set of *types* is the smallest set such that:

i. all basic types are types,

ii. if α and β are types, then $(\alpha\beta)$ is a type.

[5]Sometimes denoted as $\alpha \to \beta$, sometimes as $\beta\alpha$.

Definition 2 (Frames). A *frame* is a set of non-empty sets $\{D_\alpha \mid \alpha \text{ is a type}\}$ such that $D_t = \{0, 1\}$ and $D_{\alpha\beta} \subseteq \{f \mid f : D_\alpha \to D_\beta\}$ for all complex types $\alpha\beta$.

The sets D_α will function as the *domains* of all things of type α. Note that we do not require domains $D_{\alpha\beta}$ to consist of *all* functions of the correct type, as this would make the logic essentially higher-order and non-axiomatisable. Let us assume for each type α the existence of denumerably infinite sets of variables and non-logical constants VAR_α and CON_α. From these we can build up terms with the help of lambda abstraction, application and the identity symbol.

Definition 3 (Terms). Define, for each α, TERM_α, the set of *terms* of type α, by the following inductive definition:

i. $\text{CON}_\alpha \subseteq \text{TERM}_\alpha$; $\text{VAR}_\alpha \subseteq \text{TERM}_\alpha$;

ii. $A \in \text{TERM}_{\alpha\beta}$, $B \in \text{TERM}_\alpha \Rightarrow (AB) \in \text{TERM}_\beta$;

iii. $A \in \text{TERM}_\beta$, $x \in \text{VAR}_\alpha \Rightarrow \lambda x(A) \in \text{TERM}_{\alpha\beta}$;

iv. $A, B \in \text{TERM}_\alpha \Rightarrow (A = B) \in \text{TERM}_t$.

If $A \in \text{TERM}_\alpha$ we may indicate this by writing A_α. Terms of type t are called *formulae*. We obtain most of the usual logical signs by means of abbreviations.

Definition 4 (Abbreviations).

$\top$	abbreviates	$\lambda x_t(x_t) = \lambda x_t(x_t)$
$\forall x_\alpha \varphi$	abbreviates	$\lambda x_\alpha \varphi = \lambda x_\alpha \top$
$\bot$	abbreviates	$\forall x_t(x)$
$\neg\varphi$	abbreviates	$\varphi = \bot$
$\varphi \wedge \psi$	abbreviates	$\lambda X_{t(tt)}\ ((X\top)\top) = \lambda X_{t(tt)}\ ((X\varphi)\psi)$

The rest of the usual logical constants can be got in an obvious way.

In order to assign each term a value in a given frame, we must interpret all variables and non-logical constants in that frame. An *interpretation* function I for a frame $F = \{D_\alpha\}_\alpha$ is a function with the set of non-logical constants as its domain, such that $I(c) \in D_\alpha$ for each constant c of type α. Likewise, an *assignment* a for a frame $\{D_\alpha\}_\alpha$ is a function that has the set of all variables for its domain, such that $a(x) \in D_\alpha$ for each variable x of type α. If a is an assignment, then $a[d/x]$ is defined by $a[d/x](x) = d$ and $a[d/x](y) = a(y)$, for $y \neq x$.

A *very general model* is a tuple $\langle F, I\rangle$ consisting of a frame F and an interpretation function I for that frame. Given some very general model and an assignment, we can give each term a value.

Definition 5 (Tarski Definition). The *value* $[\![A]\!]^{M,a}$ of a term A on a very general model $M = \langle\{D_\alpha\}_\alpha, I\rangle$ under an assignment a for $\{D_\alpha\}_\alpha$ is defined as follows (to improve readability I write $[\![A]\!]$or $[\![A]\!]^a$ for $[\![A]\!]^{M,a}$):

i. $[\![c]\!] = I(c)$ if c is a constant;
$[\![x]\!] = a(x)$ if x is a variable;

ii. $[\![A_{\alpha\beta} B_\alpha]\!] = \begin{cases} [\![A]\!]([\![B]\!]), & \text{if } [\![B]\!] \in \mathrm{dom}([\![A]\!]) \\ \emptyset, & \text{otherwise} \end{cases}$

iii. $[\![\lambda x_\alpha A]\!]^a =$ the function f with domain D_α such that for all $d \in D_\alpha$: $f(a) = [\![A]\!]^{a[d/x]}$;

iv. $[\![A = B]\!] = 1$ iff $[\![A]\!] = [\![B]\!]$.

We define a (*general*) *model* to be a very general model $M = \langle\{D_\alpha\}_\alpha, I\rangle$ such that$[\![A_\alpha]\!]^{M,a} \in D_\alpha$ for every term A_α and we restrict our attention to general models. Note that in general models the second subclause of ii. does not apply (we needed it for the correctness of definition 5). The reader may verify that on general models the logical constants $\top$, $\forall$, $\bot$, $\neg$ and $\wedge$ get their usual (classical) interpretations.

The semantic notion of entailment is defined as follows.

Definition 6 (Entailment). Let $\Gamma \cup \{\varphi\}$ be a set of formulae. Γ *entails* φ, $\Gamma \models \varphi$, if, for all models M and assignments a to M, $[\![\psi]\!]^{M,a} = 1$ for all $\psi \in \Gamma$ implies $[\![\varphi]\!]^{M,a} = 1$.

Henkin [1950] has proved that it is possible to axiomatise the logic. In fact, an elegant set of four axiom schemes and one derivation rule will do the job. For details see the literature mentioned above. For the present purposes it suffices to note that β-conversion and η-conversion hold and that we can reason with $=$ and the defined constants $\top$, $\forall$, $\bot$, $\neg$ and $\wedge$ as in (many-sorted) classical predicate logic with identity.

4 Classical Logic Without Logical Omniscience

Let us apply our logic to English.[6] Since we have decided to treat the 'logical' words as non-logical constants, we can now uniformly treat all words as such. Table 1 below gives a list of all constants that we shall use in this paper, most of them named in a way that makes it easy to see which words they are supposed to formalise (the others will not directly translate words of English; their use will become apparent below). The constants in the first column of the table have types as indicated in the second column.

The idea behind the type assignment[7] is that the meaning of a sentence, a proposition, is a function that gives us a truth value in each world (and thus it is a function of type st), that the meaning of a predicate like planet is a function that gives a truth value if we feed it an individual and a world (type $e(st)$) and that an expression that expects an expression of type α should be of type $\alpha\beta$ if the result of combining it with such an expression should be of type β. So, for example, not is of type $(st)(st)$ since it expects a proposition in order to form another proposition with it; the name mary gives a sentence if it is followed by a predicate and may therefore be assigned type $(e(st))(st)$.

non-logical constants	*type*
not	$(st)(st)$
and, or, if	$(st)((st)(st))$
every, a, some, no	$(e(st))((e(st))(st))$
is	$((e(st))(st))(e(st))$
hesperus, phosphorus, mary	$(e(st))(st)$
planet, man, woman, walk, talk	$e(st)$
believe, know	$(st)(e(st))$
i	s
h, p, m	e
B, K	$e(s(st))$

Table 1: Some non-logical constants

Some easy calculation shows that the following are terms of type st.

[6]The application of type logic to the formalisation of English discussed in this section beneficed greatly from Montague [1970^a, 1970^b, 1973]. In fact we can think of it as a streamlined form of Montague semantics.

[7]Essentially this assignment was used in Montague [1970^b] and Lewis [1974].

(7) (some woman)walk

(8) (no man)talk

(9) hesperus (is (a planet))

(10) (if((some woman)walk))((no man)talk)

(11) (if((some man)talk))((no woman)walk)

(12) mary(believe((if((some woman)walk))((no man)talk))

(13) mary(believe((if((some man)talk))((no woman)walk))

Clearly, these terms bear a very close resemblance to the sentences of English that they formalise. For example, the structure of (12) is isomorphic or virtually isomorphic to the structure that most linguists would attach to the sentence 'Mary believes that if some woman is walking no man is talking'. But it should be kept in mind that these are terms of the logic and can be subject to logical manipulation.

We must of course make a connection between at least some of the non-logical constants that we have just introduced and the logical constants of the system. For example, (10) should be equivalent with (11), but as matters stand these two terms could denote two completely different (characteristic functions of) sets of worlds. Up to now, we have allowed the interpretations of the constants not, and, or, if and the like to be completely arbitrary, but it is not unreasonable to assume that at least in the actual world, which we denote with the constant i, these interpretations are standard. In order to ensure this we impose the following non-logical axioms.[8]

[8]Here and in the rest of the paper I shall let j and k be type s variables; x and y type e variables; (subscripted) P a variable of type $e(st)$; Q a type $(e(st))(st)$ variable; and p and q variables of type st. Variables are in *italic,* constants in sans serif.

A1 $\forall p((\mathsf{not}\ p)\mathsf{i} \leftrightarrow \neg p\mathsf{i})$
A2 $\forall pq(((\mathsf{and}\ p)q)\mathsf{i} \leftrightarrow (p\mathsf{i} \wedge q\mathsf{i}))$
A3 $\forall pq(((\mathsf{or}\ p)q)\mathsf{i} \leftrightarrow (p\mathsf{i} \vee q\mathsf{i}))$
A4 $\forall pq(((\mathsf{if}\ p)q)\mathsf{i} \leftrightarrow (p\mathsf{i} \rightarrow q\mathsf{i}))$
A5 $\forall P_1P_2(((\mathsf{every}\ P_1)P_2)\mathsf{i} \leftrightarrow \forall x((P_1x)\mathsf{i} \rightarrow (P_2x)\mathsf{i}))$
A6 $\forall P_1P_2(((\mathsf{a}\ P_1)P_2)\mathsf{i} \leftrightarrow \exists x((P_1x)\mathsf{i} \wedge (P_2x)\mathsf{i}))$
A7 $\forall P_1P_2(((\mathsf{some}\ P_1)P_2)\mathsf{i} \leftrightarrow \exists x((P_1x)\mathsf{i} \wedge (P_2x)\mathsf{i}))$
A8 $\forall P_1P_2(((\mathsf{no}\ P_1)P_2)\mathsf{i} \leftrightarrow \neg\exists x((P_1x)\mathsf{i} \wedge (P_2x)\mathsf{i}))$
A9 $\forall Q\forall x(((\mathsf{is}\ Q)x)\mathsf{i} \leftrightarrow (Q\lambda y\lambda j(x = y))\mathsf{i})$
A10 $\forall P((\mathsf{hesperus}\ P)\mathsf{i} \leftrightarrow (P\ \mathsf{h})\mathsf{i})$
$\forall P((\mathsf{phosphorus}\ P)\mathsf{i} \leftrightarrow (P\ \mathsf{p})\mathsf{i})$
$\forall P((\mathsf{mary}\ P)\mathsf{i} \leftrightarrow (P\ \mathsf{m})\mathsf{i})$.

These axioms tell us that an expression $\mathsf{not}\ p$ is true in the actual world i if and only if p is false in i, that $(\mathsf{and}\ p)q$ is true in i if and only if p and q are both true in i, and so on. Given these axioms many sentences get their usual truth value in the actual world. For example A7 tells us that $((\mathsf{some}\ \mathsf{woman})\mathsf{walk})\mathsf{i}$ and $\exists x((\mathsf{woman}\ x)\mathsf{i} \wedge (\mathsf{walk}\ x)\mathsf{i})$ are equivalent, A8 says that $((\mathsf{no}\ \mathsf{man})\mathsf{talk})\mathsf{i}$ is equivalent with $\neg\exists x((\mathsf{man}\ x)\mathsf{i} \wedge (\mathsf{talk}\ x)\mathsf{i})$. Axiom A10 says that there is an individual h such that the quantifier $\mathsf{hesperus}$ holds of some predicate at i if and only if that predicate holds of h at i. We can use the axioms to see that the following terms are equivalent.

$\mathsf{hesperus}\ (\mathsf{is}\ (\mathsf{a}\ \mathsf{planet}))\mathsf{i}$	
$((\mathsf{is}\ (\mathsf{a}\ \mathsf{planet}))\mathsf{h})\mathsf{i}$	(A10)
$((\mathsf{a}\ \mathsf{planet})\lambda y\lambda j(\mathsf{h} = y))\mathsf{i}$	(A9)
$\exists x((\mathsf{planet}\ x)\mathsf{i} \wedge (\lambda y\lambda j(\mathsf{h} = y)x)\mathsf{i})$	(A6)
$\exists x((\mathsf{planet}\ x)\mathsf{i} \wedge \mathsf{h} = x)$	(β-reduction twice)
$(\mathsf{planet}\ \mathsf{h})\mathsf{i}$	(predicate logic)

Let Φ be the conjunction of our finite set of axioms and let $[k/\mathsf{i}]\Phi$ be the result of substituting the type s variable k for each occurrence of i in Φ. The st term $\lambda k[k/\mathsf{i}]\Phi$ denotes the set of those worlds in which not, and, or, if etc. have their standard logical meaning. We may view the term $\lambda k[k/\mathsf{i}]\Phi$ as formalising the predicate 'is logically possible' or 'is actualizable'. The axioms thus express that the actual world is logically possible or actualizable.

We define a notion of entailment on st terms with the help of the set of axioms AX = {A1,...,A12} (A11 and A12 will be given shortly). An argument is (weakly) valid if and only if the conclusion is true in the actual world if all premises are true in the actual world, assuming that the actual world is actualizable.

Definition 7 (Weak entailment). Let $\varphi_1, \ldots, \varphi_n$ be terms of type st. We say that ψ $follows from$ $\varphi_1, \ldots, \varphi_n$ if AX, $\varphi_1\mathsf{i}, \ldots, \varphi_n\mathsf{i} \models \psi\mathsf{i}$. Terms φ and ψ of type st are called *equivalent* if ψ follows from φ and φ follows from ψ.

That terms (10) and (11) are indeed equivalent in this sense can easily be seen now. The following terms are equivalent.

$$
\begin{array}{ll}
((\mathsf{if}((\mathsf{some\ woman})\mathsf{walk}))((\mathsf{no\ man})\mathsf{talk}))\mathsf{i} & \\
((\mathsf{some\ woman})\mathsf{walk})\mathsf{i} \to ((\mathsf{no\ man})\mathsf{talk})\mathsf{i} & \text{(A4)} \\
\exists x((\mathsf{woman}\ x)\mathsf{i} \wedge (\mathsf{walk}\ x)\mathsf{i}) \to ((\mathsf{no\ man})\mathsf{talk})\mathsf{i} & \text{(A7)} \\
\exists x((\mathsf{woman}\ x)\mathsf{i} \wedge (\mathsf{walk}\ x)\mathsf{i}) \to \neg\exists x((\mathsf{man}\ x)\mathsf{i} \wedge (\mathsf{talk}\ x)\mathsf{i}) & \text{(A8)}
\end{array}
$$

In the same way we find that (11) applied to i is equivalent with

$$\exists x((\mathsf{man}\ x)\mathsf{i} \wedge (\mathsf{talk}\ x)\mathsf{i}) \to \neg\exists x((\mathsf{woman}\ x)\mathsf{i} \wedge (\mathsf{walk}\ x)\mathsf{i}),$$

and the equivalence of (10) and (11) follows with contraposition.

But if we try to apply a similar procedure to (12) and (13) the process quickly aborts. It is true that (12) applied to i with the help of A10 can be reduced to

$$((\mathsf{believe}((\mathsf{if}((\mathsf{some\ woman})\mathsf{walk}))((\mathsf{no\ man})\mathsf{talk})))\mathsf{m})\mathsf{i}$$

and that (13) applied to i can be reduced to

$$((\mathsf{believe}((\mathsf{if}((\mathsf{some\ man})\mathsf{talk}))((\mathsf{no\ woman})\mathsf{walk})))\mathsf{m})\mathsf{i}$$

but further reductions are not possible. In fact it is not difficult to find a model in which one of these formulae is true but the other is false. The reason is that it is not only the denotation in the actual world of the embedded terms (10) and (11) that matters now, but that their full meanings (i.e. denotations in all possible and impossible worlds) have to be taken into account. Not only their *Bedeutung* but also their *Sinn.* Since no two syntactically different terms have the same *Sinn,* no unwanted replacements are allowed.

Note that the solution does not commit us to a Hintikka style treatment of knowledge and belief. We have not assumed that belief is truth in all doxastic alternatives, knowledge truth in all epistemic alternatives. But we can, if we wish, make these assumptions by adopting the following two axioms.

A11 $\forall p \forall x(((\mathsf{believe}\ p)x)\mathsf{i} \leftrightarrow \forall j(((\mathsf{B}x)\mathsf{i})j \rightarrow pj))$
A12 $\forall p \forall x(((\mathsf{know}\ p)x)\mathsf{i} \leftrightarrow \forall j(((\mathsf{K}x)\mathsf{i})j \rightarrow pj))$

Here B and K are constants of type $e(s(st))$ that stand for the doxastic and epistemic alternative relations respectively. A term $((\mathsf{B}x)i)j$ can be read as: 'in world i, world j is a doxastic alternative of x' or 'in world i, world j is compatible with the beliefs of x'; $((\mathsf{K}x)i)j$ can be read as: 'in world i, world j is an epistemic alternative of x' or 'in world i, world j is compatible with the knowledge of x'.[9] If these axioms are accepted, we can reduce (12) to

$$\forall j(((\mathsf{Bm})\mathsf{i})j \rightarrow ((\mathsf{if}((\mathsf{some\ woman})\mathsf{walk}))((\mathsf{no\ man})\mathsf{talk}))j),$$

a formula that expresses that (10) holds in all Mary's doxastic alternatives. Clearly, no further reductions are possible and we can still find models such that (12) is true but (13) is false (in the actual world).

The mechanism helps us to solve some related puzzles as well. For example, (16) should not follow from (14) and (15) and it doesn't.

(14) hesperus (is phosphorus)

(15) (every man)(know(hesperus(is hesperus)))

(16) (every man)(know(phosphorus(is hesperus)))

Surely, (hesperus (is phosphorus))i reduces to $\mathsf{h} = \mathsf{p}$ after a few steps, and h and p are thus interchangeable in all contexts if (14) is accepted, but (15), if it is applied to i, only reduces to

(17) $\forall x((\mathsf{man}\ x)\mathsf{i} \rightarrow \forall j(((\mathsf{K}x)\mathsf{i})j \rightarrow (\mathsf{hesperus\ (is\ hesperus)})j)),$

and (16) applied to i can only be reduced to

[9] We may demand that B and K satisfy some axioms. The following seem a reasonable choice:
$\forall x((\mathsf{K}x)\mathsf{i})\mathsf{i}$
$\forall x \forall jk(((\mathsf{K}x)\mathsf{i})j \rightarrow (((\mathsf{K}x)j\ )k \leftrightarrow ((\mathsf{K}x)\mathsf{i})k))$
$\forall x \exists j((\mathsf{B}x)\mathsf{i})j$
$\forall x \forall jk(((\mathsf{B}x)\mathsf{i})j \rightarrow (((\mathsf{B}x)j)k \leftrightarrow ((\mathsf{B}x)\mathsf{i})k))$
$\forall x \forall j(((\mathsf{B}x)\mathsf{i})j \rightarrow ((\mathsf{K}x)\mathsf{i})j)$

(18) $\forall x((\textsf{man } x)\textsf{i} \rightarrow \forall j(((\textsf{K}x)\textsf{i})j \rightarrow (\textsf{phosphorus (is hesperus)})j))$.

Clearly, the premise $\textsf{h} = \textsf{p}$ and (17) do not entail (18).[10]

Terms (15) and (16) are the *de dicto* readings of the sentences 'Every man knows that Hesperus is Hesperus' and 'Every man knows that Phosphorus is Hesperus' respectively. Of course, we can also formalise *de re* readings, as is illustrated in (19) and (20). The reading that is formalised by (19) can be paraphrased as 'Of Hesperus, every man knows that it is Hesperus', while the other term can be paraphrased as 'Of Phosphorus, every man knows that it is Hesperus'. The reader may wish to verify that in this case the relevant entailment holds: (20) follows from (19) and (14).

(19) $\textsf{hesperus } \lambda x((\textsf{every man})(\textsf{know}((\textsf{is hesperus})x)))$

(20) $\textsf{phosphorus } \lambda x((\textsf{every man})(\textsf{know}((\textsf{is hesperus})x)))$

This possibility of quantifying-in, which the present theory shares with other semantic theories of the attitudes, distinguishes the approach from Quine's [1966] syntactic treatment. But our semantic theory is as fine-grained as any syntactic theory can be, for no two syntactically different expressions have the same meaning. The resemblance between the syntactic approach and ours is close: the syntactic theory treats the attitudes as relations between persons and syntactic expressions, we treat them as relations between persons and the meanings of those expressions. But since different expressions have different meanings, this boils down to much the same thing.

References

[1] Andrews, P.B.: 1986, *An Introduction to Mathematical Logic and Type Theory: to Truth through Proof,* Academic Press, Orlando, Florida.

[10] The present system is very weak. In fact some predictions might be *too* weak. For example, even (15) is not predicted to be valid. While I think that it might be upheld that the truth of (15) is not a *logical* truth, but a truth contingent on the properties of human belief, we are not committed to such a point of view. In order to have (15) come out valid, we may strengthen our system by adopting A9′ and A10′ below instead of axioms A9 and A10. This answers an objection that was made by Professor Paul Gochet among others.

A9′ $\forall k \forall Q \forall x(((\textsf{is } Q)x)k \leftrightarrow (Q\lambda y \lambda j(x = y))k)$

A10′ $\forall k \exists x \forall P((\textsf{hesperus } P)k \leftrightarrow (Px)k)$ etc.

[2] Barwise, J.: 1974, Axioms for Abstract Model Theory, *Annals of Mathematical Logic* **7**, 221-265.

[3] Benthem, J.F.A.K. Van, and Doets, K.: 1983, Higher-Order Logic, in Gabbay & Guenthner [1983] Vol I, 275-329.

[4] Church, A.: 1940, A Formulation of the Simple Theory of Types, *The Journal of Symbolic Logic* **5**, 56-68.

[5] Cresswell, M.J.: 1972, Intensional Logics and Logical Truth, *Journal of Philosophical Logic* **1**, 2-15.

[6] Fagin, R. and Halpern J.Y.: 1988, Belief, Awareness and Limited Reasoning, *Artificial Intelligence* **34**, 39-76.

[7] Gabbay, D. and Guenthner, F. (eds.): 1983, *Handbook of Philosophical Logic*, Reidel, Dordrecht.

[8] Gallin, D.: 1975, *Intensional and Higher-Order Modal Logic*, North-Holland, Amsterdam.

[9] Henkin, L.: 1950, Completeness in the Theory of Types, *The Journal of Symbolic Logic* **15**, 81-91.

[10] Henkin, L.: 1963, A Theory of Propositional Types, *Fundamenta Mathematicae* **52**, 323-344.

[11] Hintikka, J.: 1975, Impossible Possible Worlds Vindicated, *Journal of Philosophical Logic* **4**, 475-484.

[12] Hoek, W. Van der, and Meyer, J.-J.: 1988, *Possible Logics for Belief,* Rapport IR-170, Vrije Universiteit, Amsterdam.

[13] Levesque, H.J.: 1984, A Logic of Implicit and Explicit Belief, *Proceedings AAAI-84,* Austin, Texas, 198-202.

[14] Lewis, D.: 1974, 'Tensions, in Munitz, M.K. and Unger, P.K. (eds.), *Semantics and Philosophy,* New York University Press, New York.

[15] Mates, B.: 1950, Synonymity, reprinted in Linsky (ed.), *Semantics and the Philosophy of Language,* The University of Illinois Press, Urbana, 1952, 111-136.

[16] Montague, R.: 1970, Universal Grammar, reprinted in Montague [1974], 222-246. Montague, R.: 1973, The Proper Treatment of Quantification in Ordinary English, reprinted in Montague [1974], 247-270.

[17] Montague, R.: 1974, *Formal Philosophy*, Yale University Press, New Haven.

[18] Moore, R.C.: Propositional Attitudes and Russellian Propositions, in R. Bartsch, J.F.A.K. van Benthem and P. van Emde Boas (eds.), *Semantics and Contextual Expression, Proceedings of the Sixth Amsterdam Colloquium*, Foris, Dordrecht, 147-174.

[19] Muskens, R.A.: 1989[a], A Relational Formulation of the Theory of Types, *Linguistics and Philosophy* **12**, 325-346.

[20] Muskens, R.A.: 1989[b], Going Partial in Montague Grammar, in R. Bartsch, J.F.A.K. van Benthem and P. van Emde Boas (eds.), *Semantics and Contextual Expression, Proceedings of the Sixth Amsterdam Colloquium*, Foris, Dordrecht, 175-220.

[21] Muskens, R.A.: 1989[c], *Meaning and Partiality,* Dissertation, University of Amsterdam. Putnam, H.: 1954, Synonymity and the Analysis of Belief Sentences, *Analysis* **14**, 114-122. Quine, W.V.O.: 1966, Quantifiers and Propositional Attitudes, in *The Ways of Paradox,* New York.

[22] Rantala, V.: 1982[a], Impossible Worlds Semantics and Logical Omniscience, in I. Niiniluoto and E. Saarinen (eds.), *Intensional Logic: Theory and Applications,* Helsinki.

[23] Rantala, V.: 1982[b], Quantified Modal Logic: Non-normal Worlds and Propositional Attitudes, *Studia Logica* **41**, 41-65. Russell, B.: 1908, Mathematical Logic as Based on the Theory of Types, *American Journal of Mathematics* **30**, 222-262. Thijsse, E.: 1992, *Partial Logic and Knowledge Representation,* Dissertation, Tilburg University.

[24] Vardi, M.Y.: 1986, On Epistemic Logic and Logical Omniscience, in J.Y. Halpern (ed.), *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the 1986 Conference,* Morgan Kaufmann, Los Altos, 293-305.

[25] Wansing, H.: 1990, A General Possible Worlds Framework for Reasoning about Knowledge and Belief, *Studia Logica* **49**, 523-539.

Reinhard Muskens
Department of Linguistics
Tilburg University