Probabilistic model-theoretic semantics for *want*

Among propositional attitudes, surprisingly little semantic research has been devoted to desire verbs. The possible-worlds analysis of propositional attitudes given by Hintikka (1969), designed for verbs like *believe*, was generally assumed adequate for desire predicates as well. Stalnaker (1984) showed that such approach is inappropriate. Recent formal analyses that appeared in Asher (1987), Heim (1992), and Geurts (1998), are based on the assumption that 'wanting that φ' requires *strong preference* that φ, that is, preferring φ to $\neg\varphi$ in any possible case.

In this paper I argue that this assumption should be revised, and present a probabilistic model for sentences of the kind *d wants* φ. Instead of '*strong preference* that φ' I claim that it's the *overall expected desirability* of φ that matters.

My model is the standard possible worlds model with two additions. First, every desire report *d wants* φ is evaluated with respect to an *evaluation function* $g(w')$ that quantifies the desirability of possible worlds for *d*. The higher $g(w')$, the more desirable is w'. I argue that it is crucial to allow different evaluation functions for the same individual – a person can have contradicting attitudes toward the same thing, both to want it ant not to want it simultaneously, but according to different aspects. This explains the fact that sentences (1) and (2) can be simultaneously true self-reports of the same person. On the contrary, contradicting belief reports, like (3) and (4), cannot simultaneously describe attitudes of the same congruent person. This difference was overlooked by previous analyses.

(1) I want to play tennis now, but I have to teach.

(2) I don't want to play tennis now, because I have to teach.

(3) I believe I am playing tennis now.

(4) I don't believe I am playing tennis now.

The second component added to the model is the function $P_{d,w}(w_A = w' | [[\varphi]]_{w_A} = i)$ representing the conditional probability that the individual *d* in *w* assigns to the possibility that *w'* is the actual world (denoted w_A) , *given* that the value of φ in the actual world is *i* ($i \in \{0,1\}$).

The **truth conditions** for '*d* wants φ with respect to *g*' in the proposed model are as follows:

(5) $w \in$ [[*d* wants φ with respect to *g*]] iff

$$\sum_{w' \in W} g(w') \cdot P_{d,w}(w_A = w' | [[\varphi]]_{w_A} = 1) > \sum_{w' \in W} g(w') \cdot P_{d,w}(w_A = w' | [[\varphi]]_{w_A} = 0)$$

Intuitively, the proposed condition is that the expected desirability of the situation in the case that φ is higher than the expected desirability of the situation in the case that $\neg\varphi$.

A typical example supporting the proposed analysis is the case of a person thinking of buying a house insurance. In the most probable case, nothing happens to the house and buying the insurance would result only in the loss of the paid premium. In the very improbable, but possible case of a house being ruined, the insurance would save this person from a financial crisis. In fact, it is reasonable for a person in this situation to want to buy insurance. A sample model for considering insurance is shown in Figure 1. In this case the premium is 50, the chance for damage is 0.01, and the cost of damage, as perceived by *d*, is 100000. It is assumed that the events of buying insurance and the damage are independent.

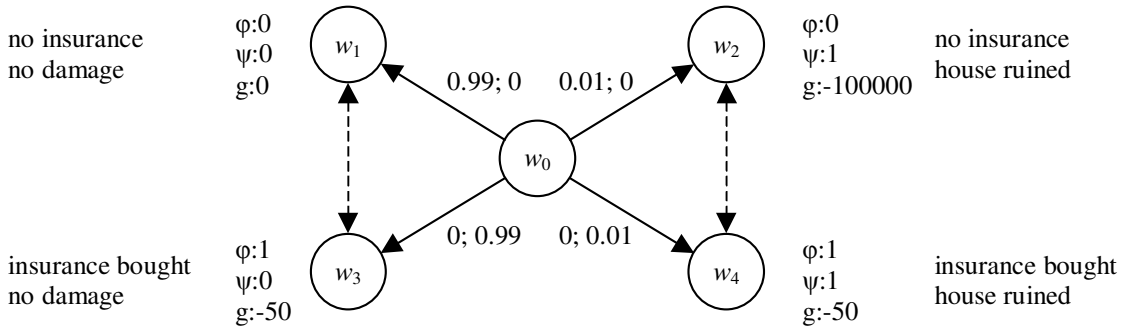

Figure 1. Model for the insurance example. φ: insurance bought. ψ: damage occurred. The numbers on the edges from w_0 to $w_{1..4}$ are conditional probabilities $P_{d,w_0}(w_A = w_i \mid [[\varphi]]_{w_A} = 0)$; $P_{d,w_0}(w_A = w_i \mid [[\varphi]]_{w_A} = 1)$. Dashed arrows show the closest alternative world given that the truth value of φ is changed.

The world w_3 in the example is the φ-alternative to w_1, and w_3 is less desirable than w_1. According to the *strong preference* based analyses, it cannot be said that the person wants to buy insurance in this case. According to my analysis, the left part of (5) is 0.99*(-50) + 0.01*(-50) = -50, the right part is 0.99*0 + 0.01*(-100000) = -1000, and -50 > -1000. Condition (5) holds, the person does want to buy insurance, and this is the correct prediction in this case.

This analysis correctly predicts that (6) is a coherent self-report, while analyses based on strong preference classify (6) as a contradiction, since the speaker reports both preferring of ¬φ in most cases and desire for φ, which is taken to mean preferring of φ in all the possible cases.

(6) I want to buy insurance. I know that most probably I'll just lose the money I'll pay for it.

My analysis is also supported by Horn's (1989, p. 326) observation that *want* is a Neg-Raising trigger, and that such triggers are almost exclusively *weakly intolerant* predicates. *Want* is indeed *weakly intolerant* according to my analysis, but *strongly intolerant* according to the others.

In addition to the examples discussed above, the proposed analysis explains many peculiarities in the inference patterns of *want* reports. My model also eliminates the need in the problematic notion of similarity between worlds, which is crucial in Heim's analysis. Furthermore, my analysis allows to demonstrate that connectedness of the predicates *want* and *good* is actually consistent with Wierzbicka's (1996, p. 51) examples meant to disprove it.

References

Asher, Nicholas (1987). A Typology for Attitude Verbs and their Anaphoric Properties. *Linguistics and Philosophy* 10, 125-197.

Geurts, Bart (1998). Presuppositions and Anaphors in Attitude Contexts. *Linguistics and Philosophy* 21, 545-601.

Heim, Irene (1992). Presupposition Projection and the Semantics of Attitude Verbs. *Journal of Semantics* 9, 183-221.

Hintikka, Jaakko (1969). Semantics for Propositional Attitudes, in J. W. Davis *et al.* (ed.), *Philosophical Logic*, Dodrecht, 21-45.

Horn, Laurence R. (1989). *A Natural History of Negation*. University of Chicago Press, Chicago.

Stalnaker, Robert C. (1984). *Inquiry*. Cambridge, MA.

Wierzbicka, Anna (1996). *Semantics: Primes and Universals*. Oxford University Press.