Reasoning about Modals

Dimka Atanassov

1 Overview

Modals provide a rich resource for investigating various aspects of reasoning about uncertainties and certainties. For instance, we all have the feeling that when using a modal expression such as *might* the speaker conveys that they are less certain about the information offered than if they had used a stronger term such as *must* or even *is*. Such intuitions are by large captured by existing semantic truth-conditional theories of modality. A question that has been much less explored is what the underlying psycholinguistic mechanism that allows us to understand modal expressions looks like. That is, most existing work focuses on the end result (i.e., the truth conditions) rather than on 'how to get there'. This research will focus on the 'getting there' part, that is, I will try to get some insight on how people reason about modals.

Varying amounts of certainty. There are may different expressions of epistemic modality-intuitively they differ by the amount of certainty they convey. For instance *must* is taken to convey more certainty than *might* is. In section 3 of this proposal I report experimental results that show evidence for different modals being associated with (significantly) different amounts of certainties ("Honest Joe" experiment). Subjects were given information presented with a different modal expression, and were then asked to make a choice based on that information and bet game points on their choice. Results from 25 subjects show significant difference between the amount of points bet when different modals were used, suggesting different amount of certainty associated with the different modal expressions. These results can give us a clear starting point as they mostly verify our intuitions about modal strength. In suggested follow up studies I propose to manipulate the experimental task in order to gain insight on how modals influence the amount of certainty associated with the information offered. In particular, it would be interesting to see whether there is a difference between the amount of certainty the speaker is *perceived* to have (by the hearer) and the amount of certainty that the hearer *decides* to attribute to the information perceived, as well as how statements with modals affect the hearer's belief state when conflicting information is present.

Lexical or derived. Is modal strength lexically determined or perhaps (in some cases) arrived at via scalar implicature (Grice [1975]; Horn [1984])? Expressions denoting degrees of certainty can be arranged on a Horn scale of the form <*possible*, *likely*, *certain*> (since *certain* entails likely but not vice versa and *likely* entails *possibly* but not vise versa). Since modals express certainty it makes sense to think about them in the context of a Horn scale. Therefore it makes sense to think about interpretation of modals as involving implicatures. For instance, if a speaker used *might* instead of *must*, this would imply that the level at which the speaker is certain is lower than what he would need in order to say the stronger *must*. In section 4 I provide some evidence for delayed processing associated with *might*, which is interpreted as a

delay in processing of the implicature associated with *might* (i.e., not *must*). I then suggest further experiments to address the issue of implicature processing.

Quantifier-like or adjective-like? There is a lot of modal-theoretic work on modals. The standard view ((Kratzer [1977]; Kratzer [1981]; Kratzer [1991])) treats them as quantifiers over possible worlds. Some recent work (Lassiter [2011]) suggests to look at modals as scalar entities (similar to gradable adjectives) and hence to directly associate them with probabilities. Both of these views capture the basic intuitions about modals in terms of truth conditions, however, the reasoning mechanism implied is fundamentally different. In the standard quantificational view modals are quantifiers over possible worlds, hence in order to tell whether a modal expression (i.e. *modal p*) is true or false, one would have to test whether the respective proposition (p) holds in some/all of the relevant possible worlds. Under the adjectival framework modal expressions are statements about probabilities, hence to verify *modal p* one would have to know the probability of p (or at least some properties of that probability, that is, whether it is lower or higher than some contextually given threshold).

We can try to get some insight on which view is more psychologically plausible by using experimental methods. For instance, there is some recent work (Larson et al. [2007]) that suggests that scalar implicatures triggered by quantifiers differ from those triggered by gradable adjectives. Hence we can look at the time course of implicatures triggered by modals and compare it to that of implicatures triggered by quantifiers and gradable adjectives. A series of experiments designed to address this issue is proposed at the last part of section 4.

In addition, in section 6 I propose an additional experimental design that will allow to directly compare modals to quantifiers. This design is inspired by the growing body of work on reasoning about quantifiers (Petroski et al. [2009]; Clark and Grossman [2007]; Troiani et al. [2007]). The idea in the quantifier-reasoning experiments is to learn something about quantificational reasoning by asking subjects to verify a statement featuring a quantifier given visual stimuli with varying complexity (i.e., varying number of items). In section 6 I propose to use a similar design with both quantifiers and modals, which would then allow us to directly compare them.

Reasoning about "must". A lot of work has been done on the 'strength' of epistemic *must* compared to *is* (Veltman [1985]; Kratzer [1991]). In recent work von Fintel and Gillies [2010] argue that *must* is not really 'weaker' but rather it signals that the speaker is basing himself on inferences. The results from the "Honest Joe" experiment in section 3 however show evidence that *must* is indeed considered weaker than *is*, as it was associated with significantly lower amounts of points bet. On the other hand in an eye tracking study comparing *must* to *is* (discussed in section 5) I found evidence for statements with *must* being interpreted as involving inferences, hence supporting the basic idea in von Fintel and Gillies [2010]. I propose that *must* is both weaker and signals that an inference was made, and in fact, *must* is weaker **because** it involves making an inference.

In section 2 I provide a short summary of the standard modal-theoretic approach to modality (Kratzer [1977]; Kratzer [1981]; Kratzer [1991]) and of a competing view that treats modals as scalar entities (Lassiter [2011]). This summary is meant to serve as background for the experimental discussion in following sections.

2 Background

It has been the standard view that modal expressions such as *must*, *might*, etc. are quantifiers over possible worlds (Kratzer [1977]; Kratzer [1981]; Kratzer [1991]). The main idea (Kripke [1963]) is that modal strength is derived via varying the quantificational force. That is, given a set of possible worlds, a modal such as *might* is interpreted as an existential quantifier, and a modal such as *must* is interpreted as a universal quantifier. For instance, *"Mary might be a millionaire"* is true if there is a world in which Mary is millionaire, and *"Mary must be a millionaire"* is true if in all possible worlds Mary is a millionaire. The set of worlds that are taken into account is determined via the context (Kratzer [1977]; Kratzer [1981]; Kratzer [1991]).

Although this view has been the prevalent one in contemporary semantics, not everyone agrees that this is the right way to treat modal expressions. For instance Lassiter [2011] proposes an alternative semantics in which epistemic modals are treated as scalar objects, similar to gradable adjectives, and their scales are associated with varying probabilities. That is, epistemic modals are not quantifiers at all, but rather probabilistic entities. For instance, *"Mary might be a millionaire"* is true if the probability of *Mary is a millionaire* is greater than some contextually given threshold. *"Mary must be a millionaire"* is also true if the probability of *Mary is a millionaire* is greater than some contextually given threshold, but this time this is a higher threshold. That is, modal strength is derived by varying the thresholds and the requirements that need to hold with respect to these thresholds.

Both of these views have theoretical appeal. A main appeal of treating modals as quantifiers over possible worlds is that this allows for a unified account of different modality flavors (epistemic, deontic, etc.). Treating modals as scalar-probabilistic expressions allows for a more elegant account of expressions such as *'p is more likely than q'*, but loses the appeal of a single unified account, since the analysis of epistemic modals cannot be generalized to deontic modality. Both of the accounts derive truth-conditions that are mostly consistent with the intuitions that we have about modals, but they do so via different mechanisms.Therefore it seems difficult to prefer one account over the other based on theoretical grounds only.

However, these two views have very different predictions on the 'reasoning' side. For instance, under a Kratzer-style quantifier approach in order to decide whether *"Mary might be a millionaire"* is true we need to go to all the relevant possible worlds and check whether Mary is a millionaire in at least one of them. If there is such a world, then the sentence is true, but if there is no such a world the sentence is false. On the gradable-adjective view things are different. *"Mary might be a millionaire"* tells us something about the chances of Mary being a millionaire, namely, that the speaker believes that there is some probability that Mary is a millionaire and that this probability is greater than some contextually given threshold. Therefore, in order to be able to evaluate *"Mary might be a millionaire"* we need to know what the probability of *"Mary is a millionaire"* is, and what the threshold is. Therefore although it might be the case that both views give very similar predictions as far as truth conditions go, they differ fundamentally in terms of the underlying reasoning procedure that they suggest. Both views on modality predict that there would be a contrast in strength between different modal expressions (with some modal expressions inducing more certainty than others), but for the quantificational account this contrast is a product of quantificational force, whereas for the adjectival account the contrast follows from different modal expressions being directly associated with different probabilities.

2.1 The standard view

In an influential series of papers (Kratzer [1977]; Kratzer [1981]; Kratzer [1991]) Kratzer proposes a theory of modal expressions as quantifiers over possible worlds, based on the account of beliefs by Hintikka [1969] and Kripke's modal logic (Kripke [1963]). She introduces two new ingredients. The first ingredient is the *conversational background*, and it is meant to capture the "in view of" meaning component of a modal. This in turn allows for a unified account of different modality flavors. For instance (1) can have many different meanings. One might utter (1) if one is making a deduction based on what one knows about Mary's whereabouts (1a), or one might utter (1) if Mary is required to be home by a certain hour by her parents (1b).

(1) Mary must be at home at 5PM.
   a. Mary must be home in view of what I know.
   b. Mary must be home in view of a certain set of requirements.

This "in view of" component is captured by the *conversational background*. Intuitively, the conversational background sets the framework (or the context) in which the modal expression is uttered. For instance, to get the interpretation in (1a) the conversational background will contain all the things known to the speaker, and for the interpretation in (1b) the conversational background will contain the set of relevant rules.

More formally the *conversational background* is is a function f such that for each $w \in W$ $f(w)$ is the set of propositions that the speaker knows/ wants/ etc. in w. $\bigcap f(w)$ is defined as the intersection of all the propositions in $f(w)$, hence, it is the set of worlds for which all the propositions in $f(w)$ are true. Then the semantic of modals is roughly defined as:

1. If N is a necessity modal then $[\![N\beta]\!]^{w,c,f} = 1$ iff $\forall v \in \bigcap f(w) : [\![\beta]\!]^{w,c,f} = 1$
2. If P is a possibility modal then $[\![P\beta]\!]^{w,c,f} = 1$ iff $\exists v \in \bigcap f(w) : [\![\beta]\!]^{w,c,f} = 1$

The second component that Kratzer defines is the *ordering source* g, a function from worlds to sets of propositions (like f). Intuitively g defines an ordering of sub-optimal scenarios, which often arise when discussing deontic modality. For instance, suppose that the following rules hold:

1. No walking on the grass allowed.
2. People who walk on the grass will be fined 5 dollars.

Suppose w_1 is a world in which Mary walks on the grass and is fined \$5, and w_2 is a world in which Mary walks on the grass and is not fined. Both are suboptimal, hence if we look only at a "perfect set of rules" both of these worlds would be marked as "equally bad". However, intuitively w_1 is better than w_2, since it contains less violations. The *ordering source* allows to capture these intuitions. By placing the two constraints above in an ordering source g we have a way of deriving that w_1 is better than w_2: the two worlds are compared in terms of 'how much of the constraints in g each world satisfies'.

Formally the function g defines a pre- order (a transitive and reflexive relation) over worlds in the following way:

$u \geqslant v$ iff $\forall p \in g(w) : v \in p \rightarrow u \in p$ (i.e., iff u satisfies every proposition in $g(w)$ that v does).

a world u is a least as good as v is if it satisfies all the propositions in g that v does.

f and g are determined pragmatically (using the context).

Then necessity and possibility modals are defined in using f and g. The simplified definition which makes use of "best worlds" is as follows:
$BEST(f(w))(g(w)) = \{v | v \in \bigcap f(w) \wedge \neg\exists v' \in \bigcap f(w) : v' > v\}$

That is $BEST(f(w))(g(w))$ is defined as the set of all worlds in the conversational background that are not dominated by any other worlds.

Then $must\phi$ is true if in all the best worlds ϕ holds, and $might\phi$ is true if for some best worlds ϕ holds:
$[\![must\phi]\!]^{M,w,g} = 1$ iff $\forall u : u \in BEST(f(w))(g(w)) \rightarrow u \in \phi$
$[\![might\phi]\!]^{M,w,g} = 1$ iff $\exists u : u \in BEST(f(w))(g(w)) \wedge u \in \phi$

Then an ordering relation on propositions is defined on the basis of the ordering on possible worlds.
Formally a proposition p is at least as good as a proposition q iff $\forall u \in q \exists v \in p : u > v$ (where $u, v \in \bigcap f(w)$)

This definition however has a problem (as noted by Yalcin [2010]). If there was a world w that was better than any other world Kratzer's theory would predict that all propositions containing w are equally good. Hence W and $\{w\}$ are equally good propositions.

To correct this Kratzer [2012] proposes a modification of comparative possibility. A proposition p is at least as good as q in a world w with respect to a conversational background f and an ordering source g iff
$\neg\exists u \in (p - q) : \forall v \in (q - p)[u > v]$

That is, when comparing two propositions we are now only comparing worlds that are not contained in both of them.

The main appeal of Kratzer's theory is that it allows to give a unified account of different flavors of modality. The difference in modal strength in turn is captured by varying the quantificational force associated with different modals, as well as the sets of worlds quantified over. However, the theoretical framework is far from simple and includes the interaction of many different ingredients: for instance, in order to account for properties of deontic modality we need to have the g function in addition to the conversational background.

2.2 Lassister's account

Lassiter [2011] proposes an alternative account in which epistemic modals are treated as scalar expressions (similar to adjectives) rather than quantifiers.

Lassiter [2011] raises several objections to treating epistemic modals as quantifiers. For instance Lassiter notes that in Kratzer's theory it is difficult to give a semantics for examples such as (2), simply because in a quantificational approach there is no straightforward way of talking about ratios and proportions.

(2) ϕ is twice as likely as ψ

Lassiter also points out is that in Kratzer's theory there are too many incompatibilities. Since $\succcurlyeq$ is a pre-order (reflexive and transitive), this leaves many epistemic comparatives and equatives undefined- any conflict of expectations leads to incomparability. In a Kratzerian model, if the model is even of moderate size about half of the comparisons would be undefined. Two propositions ϕ and ψ may be defined as incomparable even if ϕ fulfills more expectations than ϕ does. The theory has no means of comparing these two propositions unless the set of expectations fulfilled by one is a subset of the set of expectations fulfilled by the other.

Lassiter [2011] proposes an alternative semantic for modals, that treats modal expressions in a way that is similar to scalar adjectives. That is, the semantic that Lassiter proposes for modals is built on scales: modality expressions have scales associated with them, just like gradable adjectives (following the account of gradable adjectives proposed in Kennedy [2007]). For example, in Lassiter's account *possible* is a minimum standard adjective, meaning that it is associated with a scale that has a minimum associated with it. Intuitively, one would say *"It is possible that p"* if the degree of possibility associated with p is greater than that minimum (or threshold). The parallel from the gradable adjective world is adjectives like 'bent' or 'dangerous'. For example the adjective 'dangerous' has a scale with a minimum associate with it, such that one would say *"This neighborhood is is dangerous"* if the neighborhood in question has a degree of danger that is greater than the minimum on the danger scale.

More formally Lassiter proposes a scale structure $S_{epistemic}$ to capture the semantics of modal adjectives such as *possible*, *probable*, *likely* and *certain*, and then argues that this scale structure is appropriate for modal auxiliaries such as *must* and *might* as well.

Lassiter shows evidence that *possible* behaves as a minimum standard adjective (for instance, he shows that *possible* can be modified with "slightly"), hence the semantics of *possible* is associated with a minimum. Therefore the scale structure needs to have a minimum in order to account for *possible*. *Likely* and *probably* are associated with ratios, and hence to be able to account for them the scale needs to be a ratio scale (and hence $\circ$ is necessary). Finally, since *certain* is associated with a maximum, hence the scale needs to have a maximal element.

This scale is formally defined as:
$S_{epistemic} = \langle \Phi, \succcurlyeq_{epistemic}, \circ, \bot_{epistemic}, \top_{epistemic}, \rangle$, where

1. Φ is a set of propositoins
2. $\succcurlyeq_{epistemic}$ is a weak order on Φ (i.e.transitive,reflexive,and complete[1])
3. $\forall \phi \in \Phi : \bot_{epistemic} \succcurlyeq_{epistemic} \phi$ and $\phi \succcurlyeq_{epistemic} \top_{epistemic}$ (the structure has a minimum and a maximun)
4. $\langle \Phi, \succcurlyeq_{epistemic}, \circ \rangle$ is a concatenation structure[2] which is positive ($\forall x \forall y : x \circ y \succcurlyeq_{epistemic} x$), regular ($\forall x \forall y$ if $x \succ_{epistemic} y$ then $\exists z : x \succcurlyeq_{epistemic} (y \circ z)$), and Archimedian (if $a \succcurlyeq_{epistemic} b$ then for all c and d there is a positive integer n such that $na \circ c \succcurlyeq_{epistemic} nb \circ d$. na is defined inductively as $1a = a, (n+1)a = na \circ a$).

Then measure functions are defined as functions from this algebraic structure to the real numbers. Admissible measure functions are functions that preserve the properties of the algebraic structure (i.e., preserve the ordering, positivity, etc).

[1] Meaning that every two items can be compared, which is not the case with partial orders

[2] $\circ$, the concatenation relation is a ternary relation, which can be thought of as a a function $X \times X \to X$

Skipping some formal details, Lassiter shows that this scale structure is mathematically well behaved (equivalent to a representation by finitely additive probability). In addition, all admissible measure functions μ are identical up to multiplication by a scalar (i.e., for every μ and μ' there is some $r \in R^+$ for which $\mu'(x) = r\mu(x)$). In fact, if we decide that the maximim is mapped to 1 ($\mu(m) = \mu(W) = 1$) we are left with only one admissible measure function, which Lassiter calls *prob*.

Hence $prob(\phi) > prob(\psi)$ means that for all $S_{epistemic}$ admissible μ $\mu(\phi) > \mu(\psi)$. The important part is that there is nothing special about *prob*, since it is isomorphic to all other $S_{epistemic}$ admissible μ. It is chosen because of convention.

The proposed meaning for the above modal expression is then as follows:

- $[\![\phi \text{ is possible}]\!]^{M,w,g} = 1$ iff $\mu_{epistemic}(\phi) > \mu_{epistemic}(\bot_{epistemic})$, which translates to $prob(\phi) > 0$
- $[\![\phi \text{ is likely/probable}]\!]^{M,w,g} = 1$ iff $\mu_{epistemic}(\phi) > \mu_{epistemic}(\theta_{epistemic})$ with $\theta_{epistemic}$ some vague threshold, which translates to $prob(\phi) > \theta_{epistemic}$.
- $[\![\phi \text{ is certain}]\!]^{M,w,g} = 1$ iff $\mu(\phi) = \mu(\top_{epistemic})$ which translates to $prob(\phi) = 1$

Lassiter then proposes to extend the scale for the auxiliary modals in the following way:

- $[\![must(\phi)]\!]^{M,w,g} = 1$ iff $prob(\phi) \geqslant 1 - \alpha$ where α is contextually given (Swanson 2006). (Lassister does not supply a scale definition, but it is possible to build one)
- $[\![might(\phi)]\!]^{M,w,g} = 1$ iff $prob(\phi) > \alpha$, where α is contextually given (again, no scale definition is given)
- $[\![should/ought]\!]^{M,w,g} = 1$ iff $prob(\phi) > \theta_{epistemic}$, $\theta_{epistemic}$ is the same as in the definition of likely and probable[3].

The probabilities then are directly defined on propositions, where propositions are sets of worlds as in the standard theory[4]. The difference from the standard theory is that modals are not quantifiers over worlds, but rather scalar entities. The scale over which they are defined is an algebraic structure, which is isomorphic to numerical probability.

2.3 Summary

The adjectival account allows for a much more fine-grained distinction between modality strength: it allows one to talk not only about the edges of the scale (possible vs. necessary) but also about all the intermediate values (it becomes easy to give a semantics for something like *"There is a 50% chance of rain tomorrow"*). The quantificational account is more constrained-there is either an existential or a universal quantifier, and nothing much in between. To say something more fine-grained becomes increasingly difficult and usually requires sophisticated manipulations of the sets of worlds being quantified over. One might have been satisfied with

[3]It is worth noting that it is somewhat surprising to associate *should* and *ought* with the same threshold as *probably* or *likely*. Lassiter does not give motivation for this move.

[4]Lassiter also gives a variation in which probabilities are defined over possible worlds, which in turn allows to define probabilities on propositions via probabilities on possible worlds.

the quantificational view if possibility and necessity were all there ever was. However, even if we exclude expressions such as *"p is more likely than q"* or *"There is 50% chance that p"* from what we are willing to call 'modals' there still seem to be a lot of remaining modal expressions that are not 'about the edges'. For instance, we do get the feeling that although expressing necessity, *should* and *aught to* are weaker than *must*. While it is not impossible to account for this intuition in the quantificational framework by adding more ingredients (von Fintel and Iatridou [2006]), it becomes much more straightforward to do so once we have a scale (simply associate the weaker necessity modals with a lower threshold).

It is worth noting that a main disadvantage of the adjectival account is that it works only for epistemic modals. It does not generalize to deontic modality, hence a separate account is needed for deontic modals.

The main point of difference between the two views is the implications they have about the way people interpret modals. Rather than encoding modal strength via different quantificational force, the adjectival account encodes it by varying the certainty thresholds associated with modals. Therefore under the adjectival view being able to tell whether *must p* is true requires one to know what the contextually given threshold is, and the ability to assess the probability of *p*. This is much like being able to tell whether *Mary is tall* is true: to be able to do that one needs to know what the contextual threshold of *tall* is (i.e., the minimal amount of height that one must have to be considered 'tall') and what Mary's height is. This is different from the quantificational account. Under the quantificational account *must p* would be true if *p* was true in all relevant possible worlds. Therefore being able to tell whether *must p* is true requires both the knowledge of what the relevant worlds are and a procedure for checking whether or not *p* is true in each and every one of them.

It may be the case that these two views are 'equivalent' in terms of truth conditions, in the sense that one could derive the right truth conditions for modals in both of them. However, the reasoning process that these two views suggest is very different. Therefore, by looking at the reasoning process we may be able to derive some insight about which view is more psychologically plausible. Moreover, these two theoretical frameworks provide us with the types of psycholinguistic questions we could ask.

3 Modal strength: the "Honest Joe" experiment

English epistemic modals (and epistemic modals in most languages) differ from each other in terms of the strength of the statement made by using them. For example, almost everyone agrees that *"It might be raining"* is weaker than *"It must be raining"*. The first statement is understood as conveying less certainty that it is raining than the second one. Agreement tends to be less universal however when comparing *"Jane must be in her office"* to *"Jane should be in her office"* or *"Jane has to/ought to be in her office"*. We get the intuition that *must* is 'stronger' than *should* or *has to/aught to*, but there is not much theoretical agreement about how exactly those modal expressions differ, and on what theoretical grounds. And what about *"Jane must be in her office"* compared to *"Jane is in her office"*? A lot of work (von Fintel and Gillies [2010]; Kratzer [1991]; Veltman [1985]) has been done specifically when comparing the last pair of sentences, and various theoretical claims have been made.

Intuitively it seems that some version of the following ranking exists:
is $\geqslant$ must, has to $\geqslant$ should, ought to $>$ might, could

A lot of theoretical work has been done on comparing specific pairs of expressions on this ranking (von Fintel and Iatridou [2006]; von Fintel and Gillies [2010]; von Fintel and Gillies [2007]; Kratzer [1991]).

However, intuitions about this ranking have not been systematically empirically tested. In the remainder of this section I will discuss an experiment that can provide some systematic evidence about the ranking of these expressions. In the experiment people are asked to answer questions and 'bet' on their answers, based on information they are given by a fictional character ("Honest Joe"). The information that Honest Joe gives them is presented using different modal expressions. I assume a correlation between the amount of certainty that the modal expression conveys (strength) and the amount of points bet: stronger modals should solicit higher bets, and weaker modals should invite lower bets. In particular, if a certain type of modal is universally perceived as weaker than another modal, there should be a difference in the amount of points bet.

3.1 Honest Joe

This study is designed to test how sensitive people are to information presented with different modals in terms of perceived certainty and reliability of the information presented. Each time subjects were presented with scenes containing partial information and with additional information presented using different modals. They then were asked to answer a question about the scenes based on that information. Crucially, to answer the question they had to rely on the information that was presented with the modal statement, and they also had to "bet" on their answer, using game points. The "bet" that subjects placed on their answer each time (a number between 1 and 100) can tell us something about how certain subjects were about their answer, and in turn, how reliable the information presented with the modal was perceived to be.

Materials and Design.
Subjects were told that each time they were going to see a scene with three containers, and each one will contain one of three objects. They would be able to see where one of the objects is, and they would be asked to guess where one of the other two objects is. They were also told that they would be playing for points, and that right guesses would earn them game points, while wrong guesses would result in loss of game points, in the following way:

1. They start the game with a set amount of points (6000)
2. For each scene they are given a choice of the amount of points they want to play for, they can choose any amount between 1 and 100 points.
3. If their guess is correct, they double the amount they chose to play for, but if the guess is wrong they lose the game points they risked. For instance, if they chose to play for 10 points, and they got it right, they earn 20 points (and also keep the 10 points they risked), but if they guess wrong, they lose the 10 points they risked.
4. Their score would be revealed to them at the end of the experiment, but would have no consequences outside of the experiment.

Then a fictional character, "Honest Joe" was introduced. Subjects were told that:
"Honest Joe knows something about what some of the containers contain, but he does not

know everything. Honest Joe is completely honest, and he will never purposefully lie to you. For each scene Honest Joe will try his best to help you figure out where the various items are."

The fictional character, Honest Joe would then present a "clue" using one of six auxiliaries, five modals *must*, *might*, *could*, *should*, *has to*, and *is*. Each subject saw each scene only once (with one of the above auxiliaries), but the auxiliary used with each scene varied between subjects, hence there was a total of six lists. There were 36 target trials, such that each subject saw six trials with each one of the six modals. An example of a target trial is shown in figure 1.

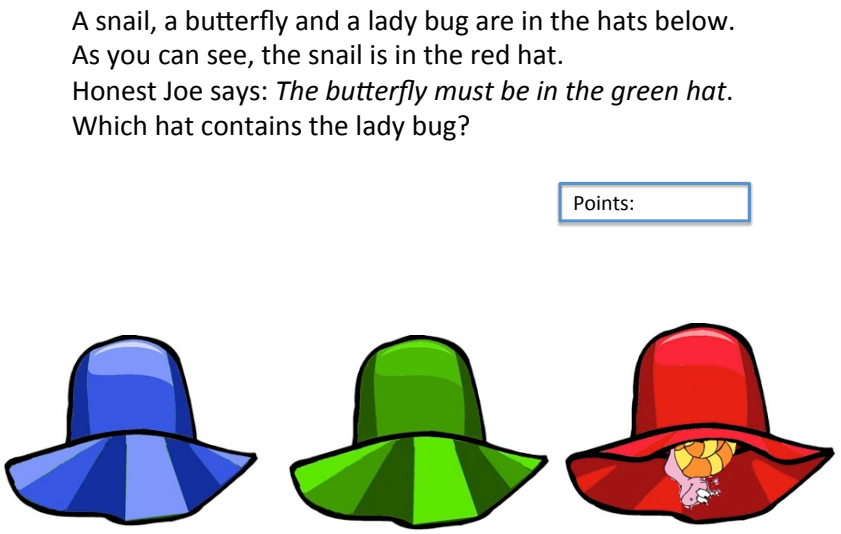

Figure 1: Example of a target trial

There were 24 fillers, for half of them no information was presented (Honest Joe was silent), for another six Honest Joe used "I think that.." instead of a modal, and for another six Honest Joe used "I know that..."

The suggestions Honest Joe provided were correct half of the time across the board. That is, they did not match our intuitions about certain expressions conveying higher chances of the information being correct. In the current version of the experiment this could not have any effect on subjects' behavior, since they did not see their score until the end of the experiment. However, in follow-up studies I intend to give subjects feedback throughout the experiment (in the middle of the experiment and after each 10 trials). It is important to keep the probability of Honest Joe's clues being correct independent of the modal expression (or at least not matching our intuitions), in order to avoid creating the effect we are looking for. Keeping the correctness of the information provided by Honest Joe at chance avoids this issue from occurring for the follow up studies (in which feedback will be provided throughout the experiment).

Results.

27 undergraduate students participated in this study in exchange for extra credit. The stimuli were presented using EPrime2 software. Two participants were excluded from the study (one was excluded due to misinterpretation of the experimental task, and another subject was excluded because he bet the maximum 100 points on all trials).

The dependent variable was the amount of points bet. In addition, *accuracy* measures whether subjects followed Honest Joe's suggestion (i.e., the accuracy was 1 for a trial in which the subject based his selection in accordance with what Honest Joe suggested, and 0 otherwise).

A summary of the results is shown in figure 2[5].

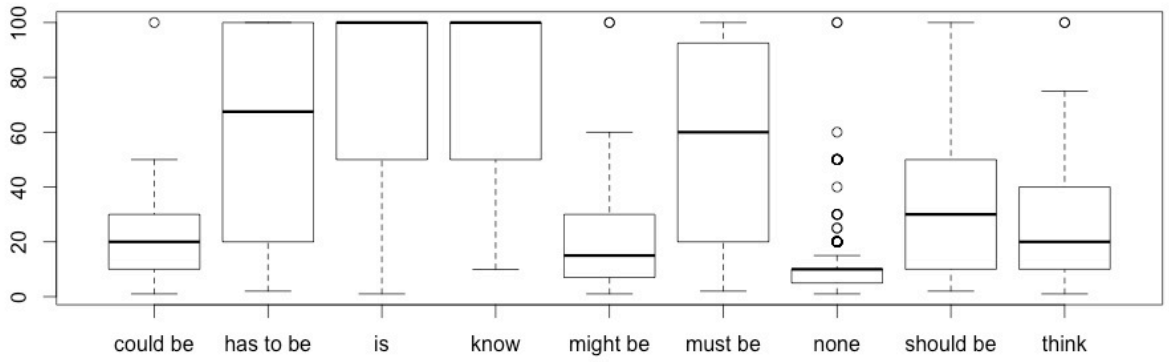

Figure 2: Boxplot showing the amount of points bet varying by the type of modal used. In the"none" condition Honest Joe did not say anything.

Although there is large variability between subjects of the amount of points they were willing to bet (due to the fact that some people are more risk averse than others) the results point to a clear pattern:

- "Is" and "know" are associated with the biggest amount of points, followed by "has to be"
- "has to be" is associated with slightly more points than "must", but the difference is not statistically significant
- "should be" is located below "must"
- "could be" and "might be" are approximately the same, and both are below "should"
- When the Honest Joe character did not offer any information (the "none" items) the least amount of points was bet.

The mean amount of points bet (as well as mean accuracy, where accuracy encodes whether the subject followed Honest Joe's suggestion) are given in table 1 below.

To see which ones of the paris of modals were significantly different from each other I conducted pairwise comparison between several different pairs of modals using a mixed effects model[6].

The results are summarized in table 2 below.

[5]Trials with 0 accuracy were taken out, except for the 'none' items, in which Honest Joe made no suggestion).

[6]Again, taking out trials with 0 accuracy, except for 'none' items

modal	Bet	Accuracy
could be	22.46667	0.7600000
has to be	59.04000	0.9066667
is	76.93333	0.9266667
know	75.46000	0.9400000
might be	20.96000	0.7800000
must be	56.18000	0.9000000
should be	39.96667	0.8666667
think	27.25333	0.9000000
none	11.53000	

Table 1: Mean amount of points and mean accuracy.

These results can be unpacked into the following empirically motivated Horn scale (Horn [1984]; Horn [2004])[7]:

$< is, \{must, has - to\}, \{should\}, \{might, could\} >$

The leftmost *is* is most informative, followed by *must* and *has to*, and so on. Thus if for example a speaker utters *might* this would result in a scalar implicature that it is not the case that a stronger statement (for instance one formulated with *should* or *must*) holds, or that there is not enough evidence to support that stronger statement.

The contrast between *must* and *is* forms an interesting case of study. In the scale above I have placed *must* to the right of *is*, however it is not clear that the contrast between the two is a matter of informativeness. In recent work von Fintel and Gillies [2010] propose that the difference between *must* and *is* is really a matter of *must* signaling that the information offered is based on making an inference, and hence *must* is not weaker than *is*. However, my current results show a clear difference between the two, with *must* being taken as significantly less reliable than *is*. A possible explanation of this data with respect to the theoretical analysis in von Fintel and Gillies [2010] is to attribute the contrast between *is* and *must* to inference making. This can be done by arguing that information based on reasoning is taken as less reliable. Since *must* is based on reasoning, it is perceived as weaker than *is*. Hence the difference between the two is not simply a matter of informativeness. (In section 5 below I offer a more extensive discussion of *must* vs. *is*).

Hence for the time being I will take *is* out of the scale, since the distinction between it and *must* is not a simple matter of informativeness. This leaves us with:

$< \{must, has - to\}, \{should\}, \{might, could\} >$

3.2 Discussion and follow-up studies

The above results show a clear pattern of the following 'ranking' between the various modal expressions, with *is* and *know* associated with the highest bet, followed by *has to* and *must* (with no significant difference between the latter). *Should* is placed in the middle, lower than

[7]expressions in curly bracket denote that they are treated as being equally informative, based on the experimental results

	modal-pair	p -value
could	could-might	0.9492
	could-think	0.0022 (**)
	could-is	<.001 (***)
	could-should	<.001 (***)
	could-must	<.001 (***)
	could-has to	<.001 (***)
	could-none	<.001 (***)
has to	has to-must	0.216
	has to- is	<.001 (***)
	has to- know	<.001 (***)
	has to- should	<.001 (***)
might	might-think	0.0038 (**)
	might-is	<.001 (***)
	might-should	<.001 (***)
	might-must	<.001 (***)
	might-has to	<.001 (***)
	might-none	<.001 (***)
must	mist-is	<.001 (***)
	must-know	<.001 (***)
	must-should	<.001 (***)

Table 2: Pairwise comparison between modals.

must and *has to*, but higher than *could* and *might*. Finally, *might* and *could* still elicited higher bets than when there was no information provided.

It is worth noting that although the amount of points bet gives an indication of 'certainty', as higher bets are associated with higher certainty, it does not map directly onto a certainty scale (i.e., if a subject bet 70 points out of 100 then that by itself does not mean that that the subject was 70% certain of his answer). This is because other factors, such as risk averseness come into the play when people are being asked to 'bet', or even more generally to make decisions (Baron [2004]; Baron and Hershey [1988]; Lopes [1996]; Birnbaum and Chavez [1997]; Tversky and Kahneman [1992]). However, the results reported here are a clear indication of a difference in behavior associated with a difference in the underlying belief system. If people consistently bet more points when the modal was *must* than *might*, it is reasonable to argue that this is because they were more certain when the information was presented to them with *must* than when it was presented to them with *might*.

Feedback. In the present study the score was only given to subjects once- at the end of the game. This was done in order to not give them opportunities to alter their behavior based on how well they were doing. However, it would be interesting to see in which way providing some 'point feedback' could lead to change in behavior, and whether (and to what extent) this would have an effect on the amounts bet. Two modifications of the present study are thus proposed. In the first modification subjects will be given feedback twice: once in the middle of the experiment and once again at the end. In the second modification people will

be presented with feedback after each 10 trials[8]. In the first modification the first and second half of the experiment will be compared, and in the second each of the 6 sub-parts will be compared to each other. The prediction is that people in fact will change their behavior in the following way: greater amount of wins would result in higher bets for following sections, but the relation between modals will stay the same (i.e., people will just bet more across the board).

Certainty gaps. The distinction between different expressions of modality was so far examined only to the extent that they induce certainty in the hearer (i..e, by examining the amount of points bet). But we can also look at an additional parameter: whether there is a difference between the amount of certainty a speaker is perceived to have when uttering *modal* p and the amount of certainty actually attributed to p by the hearer, as well as whether this difference varies across different expressions of modality.

Discourse participants are not only 'hearers' (recipients of information) but also 'speakers' (producers of information). Therefore a modal expression would be associated for each hearer with some amount of certainty that they need to have in order to produce a statement with it, which would translate to the amount of certainty they would assume another speaker has in order to produce the same statement. Suppose we have two discourse participants H and S. S utters $modal_i$ p. Suppose that H would utter $modal_i$ p if he was about 20% certain that p. Thus it is reasonable to assume that H would deduce that S is 20% certain that p as well. However, that does not mean that H would also adjust his probability of p to 20%. There are many reasons why H might not want to take the statement of S at face value: H may not trust S, or he may not trust the competency of S, or he might hold some relevant information that S does not have. However, it may be the case that the modal itself has an influence on how much H wants to 'adopt' the belief state of S. For example, it may be the case that non modal expressions are 'taken at face value' more than expressions of strong modality, which in turn are taken 'at face value' more than expressions of weak modality.

Let *perceived belief* be the amount of certainty that a hearer thinks the speaker has when uttering a modal statement, and *accepted belief* be the amount of certainty that the hearer adopts. Then define *certainty gap* as the difference between the two. In a follow up study I want to study these certainty gaps, and see to what extent they are influenced by the expression of modality (if at all).

In the study I will use the same stimuli I used in the original experiment, but the experimental task will be slightly modified. Instead of asking subjects to place bets, subjects will be directly asked for their judgements on two questions of the following type (they will be asked to indicate a number on a scale, for instance, between 1 and 7):

1. How certain are you that object X is in container Y?

2. How certain do you think that Honest Joe is that object X is in container Y?

The design will be within subjects in the sense that each subject will see only one of the above question associated with a given scene (hence instead of having 6 conditions there will now be 2x6 conditions: *modal* x *type of question*). Each subject will see both questions, but only one type of question per scene.

[8]Feedback cannot be provided after each single trial, because then this would be a direct feedback on whether Honest Joe's information was correct.

If different modal expressions induce different certainty gaps I expect to see interaction between type of modal expression used and type of question asked.

A variation in *certainty gaps* can also serve as an argument for a scalar account, since such variation can easily and elegantly be explained on a scale. The quantificational Kratzer -style account will however have difficulty explaining such an effect. If all that *might* p does is express that there is some best world in which p holds, then a difference between certainties is not predicted, since the hearer can either accept or reject what the speaker said. That is, a hearer will either accept that there is some best world in which p holds or reject that altogether. The theory simply does not have a mechanism for allowing a hearer to manipulate the level of certainty of the information he is presented with: there are no means by which a hearer could adjust his beliefs 'halfway' between what he initially believed and what he believes that the speaker believes- it is "all or nothing" in the sense that as a hearer you can either accept or reject the information you are presented with (you are not allowed to "add your own grain of salt").

If there is such a gap effect, one could argue that it is really a pragmatic effect and therefore not part of the core meaning. Hence one could argue that we should not be worried if the quantificational account cannot explain it. However, then we would have to add some pragmatic machinery on top of the quantificational semantics in order to explain the effect, and this machinery will eventually have to be about manipulating probabilities.

The above question might be taken even further- we could ask how expressions of modality are perceived when there is some conflicting information present. For instance, suppose I say *It must be raining outside.* If you have heard today's weather forecast and know that there was no rain expected, you would probably be a lot less likely to believe me than if (in the same situation) I said *It's raining.* If the first *certainty-gap* follow-up study produces interesting results, a second follow-up which would examine certainty gaps in the context of conflicting information will be launched.

4 Implicatures

The notion of implicatures was first introduced by Grice [1975], as those aspects of what is conveyed that go beyond the truth-conditional meanings. Generalized conversational implicatures in particular arise under normal circumstances unless there is something in the context that prevents them from doing so. Scalar implicatures are a type of generalized conversational implicatures which arise when a speaker uses a non-maximal value on some salient scale. The implicature is that stronger (greater) values are either false or unknown. This definition lumps together rankings, quantifiers, modals, cardinals and gradable adjectives. The following are examples of quantificational, modal and gradable adjective scalar implicatures:

(3) What happened to the chocolate cookies?
Sam ate some of them.
Implicature: Sam did not eat all of the cookies

(4) Do you know where Katie is?
She might be in her office.
Implicature: I am not certain enough to use a stronger statement, such as one with *must*.

(5) Is Pete overweight?
He's chubby.
Implicature: Pete is not fat.

It has generally been assumed that different scalar implicature triggers are not very different from each other and trigger the same reasoning. However, in recent work Larson et al. [2007] provide some evidence that people treat different scalar implicatures differently. They presented subjects with pieces of conversation of the following type:

(6) Irene: How much cake did Gus eat at his sister's birthday party?
Sam: He ate most of the cake.
FACT: By himself, Gus ate his sister's entire birthday cake

Subjects were then told to evaluate the underlined utterance (as 'True' or 'False') based on the fact and according to three different types of instructions, manipulating whether the sentence should be taken literarily or not.

The authors report many interesting results, including significant difference in responses between the different types of instructions, as well as difference between different types of stimuli (i.e., generalized implicatures, contradictions, entailments and necessary contextual entailments). Interestingly, they also found differences between the different types of scalar implicatures.

The authors looked at the amount of 'False' responses and report the following median (of 'False' responses) for the scalar implicature types:

1. Adjectives: 12%
2. quantifiers and modals: 33%
3. Rankings: 49%
4. cardinals: 75%

That is, people were a lot more willing to accept violations of adjectival implicatures than violations cardinal implicatures or a quantificational ones, for instance. This leads us to think that there is more to implicature trigger type than was standardly assumed. In particular, this study points towards a difference between adjectival scalar implicatures and quantificational scalar implicatures. The authors lump quantifiers and modals together, but even so, these results seem to be pointing towards a difference in interpretation between quantificational and adjectival implicature triggers.

Suppose that modals are in fact closer to adjectives. Then the fact that even when they are put together with quantifiers there still is a big difference between the two groups in terms of rates of rejection points towards there being a significant difference between the quantifier and the adjective group.

Suppose now that the grouping of modals with quantifiers is on the right track. Then the results can be taken at face value- i.e.- there is a difference between quantificational and adjectival implicature triggers.

The next step would be to separate the modals from quantifiers and see how they would behave as a category by themselves (and whether it would make sense to lump them with the

adjectives or perhaps with the quantifiers after all). If they behave on a par with adjectives, then we have evidence for Lassiter's account. If they behave more like quantifiers, then we have evidence for the quantificational account.

If there is a difference in rejection rates between quantifier triggered and adjective triggered implicatures there should be a difference in processing as well. One could then look at the time course of the implicatures triggered by quantifiers and adjectives, and compare that to the time course of implicatures triggered by modals. In what follows I will discuss this option in some more detail. I will present results that suggest delayed implicature processing for *might* and suggest a few new studies.

One could also do a similar study to the one reported here, but concentrating specifically on three categories: quantifiers, modals and adjectives.

4.1 *Must* vs. *Might*: delayed implicature processing

In an eye tracking study I examined the time course of the reasoning involved in *might* (compared to *must*). The study showed a delay of 800ms associated with *might*, thus supporting the claims that (a) there is an implicature involved in interpreting *might* (rather than just lexical meaning) and (b) this implicature is delayed.

Materials and Design. I employed the visual world paradigm, incorporated into a game of guessing with a confederate, in order to examine the processing of *might* and *must*. The experiment's design was inspired by the design in Brown-Schmidt et al. [2008] and Keysar et al. [2000]. The participant and the confederate played a game of guessing on a computer screen. Each saw a 3x3 display of colored shapes in which some of the shapes were occluded for one of the participants (the guesser), and he had to guess what they were, aided by two simple rules. The other participamt (the verifier) could see all shapes and had to respond to the guessers guesses by marking them as correct or incorrect. The subject and confederate took turns in playing the guesser and verifier: at first the subject was making guesses, while the confederate verified, and in the second part the subject and the confederate switched roles. All critical trials took place when the confederate was making the guesses. The first part (in which the subjects guessed) was done to familiarize subjects with the rules of the game, as well as to make the game seem more believable. While for filler trials the confederate made spontaneous utterances, for the target trials the confederate's screen contained written instructions on what to say. When the confederate was guessing the subject knew which parts of the scene were privileged, and could only be seen by him.

For the duration of the experiment, the guessers' guesses were recorded via a microphone, and the verifiers eye gaze was recorded via an eye tracker. Only the second part of the experiment, during which the confederate guessed, contained the critical, target trials. During critical trials these guesses were phrased as guesses using either *must* or *might*, depending on whether the confederate could infer his guess with certainty.

Participants were told that there are four possible shape types: hearts, circles, triangles and squares, and four possible shape colors: red, green, blue and black. Participants were also told that the shapes will always be arranged to follow two simple rules: the first rule was that in each row either all shapes will have the same type or all shapes will have different types; the second rule was that in each column either all shapes will have the same color or all will have different colors (see figure 3). All displays indeed followed these rules. An example display is shown in figure 3: the left side shows what the verifier (subject for target trials)

sees, and the right side shows the screen for the guesser (confederate for target trials). The confederate screen also contained instructions for target trials. As can be seen in this figure, all shapes in the top row are squares, and in the middle and bottom rows all shapes are of different types (a heart, a circle and a triangle in the middle row, and a square, a heart and a triangle in the bottom row). As for the columns, in the leftmost and rightmost column all shapes have the same color, red, and in the middle column all shapes have a different color: blue, green and black.

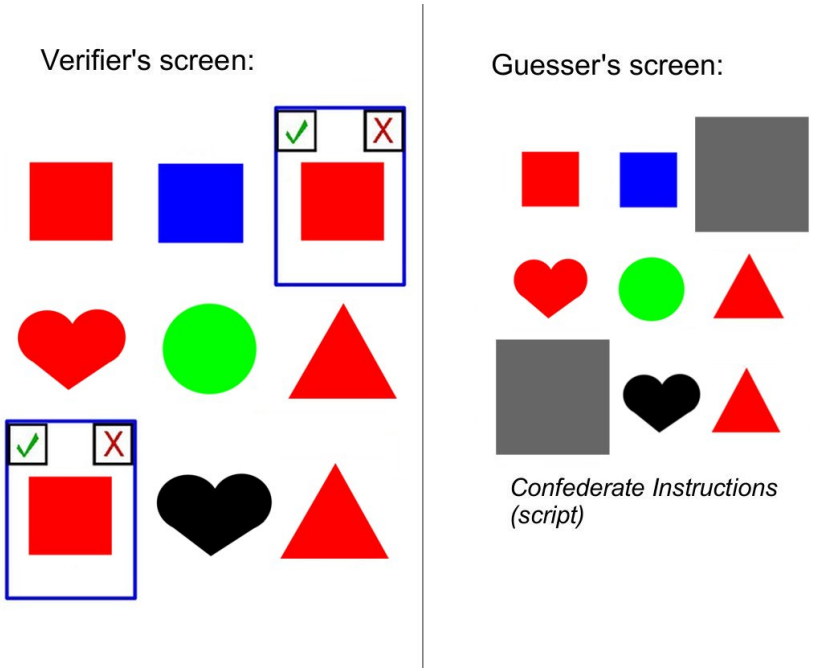

Figure 3: Must-Might: example of a target trial. confederate sees right side and subject left side.

In each trial, the subject and confederate were given five seconds to look at the screen, after which a sound was played and the person playing the guesser could start making guesses. This was done to allow the subject to familiarize himself with the scene. In addition, in order to stimulate active, rather than passive listening to the confederates guesses, the subjects were told that they should try and respond to the guesses as quickly as possible.

Target items consisted of 12 sets of partially scripted utterances, corresponding to 12 scenes. The confederate read off his screen the scripted utterances while pretending that these were spontaneous guesses. The utterances referred to one of two hidden shapes in each scene. In the example above (figure 3), the confederate uttered one of the following two sentences:

1. There **must be** a red square located in the upper right.
2. There **might be** a red square located in the bottom left.

Whenever *must* was used the location referred to a shape that the confederate could guess with certainty (in figure 3 the shape on the upper right has to be a square, because the two other shapes in the same row are squares, and it has to be red, because all other shapes in the same column are red, therefore, it can be guessed with certainty). Whenever *might* was used, the location referred to a shape that could not be guessed with certainty (in figure 3 it can be inferred that shape on the bottom left is red, because all other shapes in the same

column are red, but the shape type cannot be inferred with certainty, as it can be either a square or a circle). Whenever *must* was used the guess made was correct. For half of the cases in which *might* was used the guess made was correct, and for the other half the guess was not correct (although it was possible given the rules).

The target sentence was preceded by an introductory sentence of the form *Hmm/ Lets see, In the [row column] there is a [color shape].* This sentence always referred to a non hidden shape, that was relevant to making the guess. That shape was always of the same type and color as the one that could be guessed with certainty. In the example above, the introductory sentence was *Hmm, in the upper left there is a red square.* [9]

Participants. 14 undergraduate students participated in this study for course credit. All participants were native English speakers and undergraduates at the University of Pennsylvania. Two participants were dropped from the analyses reported below, due to high track loss percentage.

Apparatus. The images were presented on a 17 Samsung screen with 1680x1050 resolution, which was divided into two halves, such that the participant viewed one half and the confederate the other half. The division was made using a partition that did not allow the subject to view the confederates screen and vice versa. The audio recording was made using a headset with a microphone. The onset of the recording was synchronized with the display of the scenes. Right eye gaze was recorded using an Eyelink 1000 eye tracker on a desktop mount at a sampling rate of 1kHz (resampled offline to 100Hz).

Results The critical period was defined to be the time between 500 milliseconds before the onset of the modal and up until 2 seconds after the onset of the row indicator. The target was taken to be the shape that could be guessed with certainty (corresponding to the "must"), and the competitor the shape that could not be guessed with certainty (corresponding to the "might"). There are two conditions: *must*, corresponding to when the subject heard "must", and *might*, corresponding to when the subject heard "might". Target advantage score was computed as looks to the target minus looks to the competitor. This measure is appropriate since it shows more clearly the preference of the target over the competitor.

Figure 4 shows the target advantage from 500ms before the onset of the modal and until 2 seconds after the onset of the column word, for the two conditions: *must* and *might*. This figure shows that around 1000ms after the modal onset there is a clear divergence between the *must* and the *might* conditions. At 2000ms (2 seconds) after modal onset the divergence becomes even bigger. This graph then can serve as a sanity check, as when the location is revealed, the Target Advantage score for the competitor vs. target diverge appropriately.

The ambiguity period was defined as starting from the modal onset (*must* or *might*), and until 200ms after the onset of the row locator. Figure 5 shows the target advantage for this period, for items for which the guess made was correct (half of the items)[10].

[9] For the first three subjects the introductory sentence that was used was slightly different: the shape and color were specified prior to the location, i.e confederate uttered: *Hmm, there is a red square on the upper left.* This was changed for later subjects to the version above, in order to avoid subjects mistakenly responding to that sentence as the guess.

[10] An effect of condition was only present in those items that had correct *might* guesses. When the guess was wrong for the *might* condition, the shape type that was guessed matched the one in the target position (color matched both the target and the competitor). That is, the shape that could be guessed with certainty (the target shape) matched the one that the confederate uttered. That shape was also occluded, and thus also in the participants privileged ground. Thus it may be the case that when the guess was wrong the subjects simply looked at the shape that matched the one they heard (which happened to be the must shape) and was

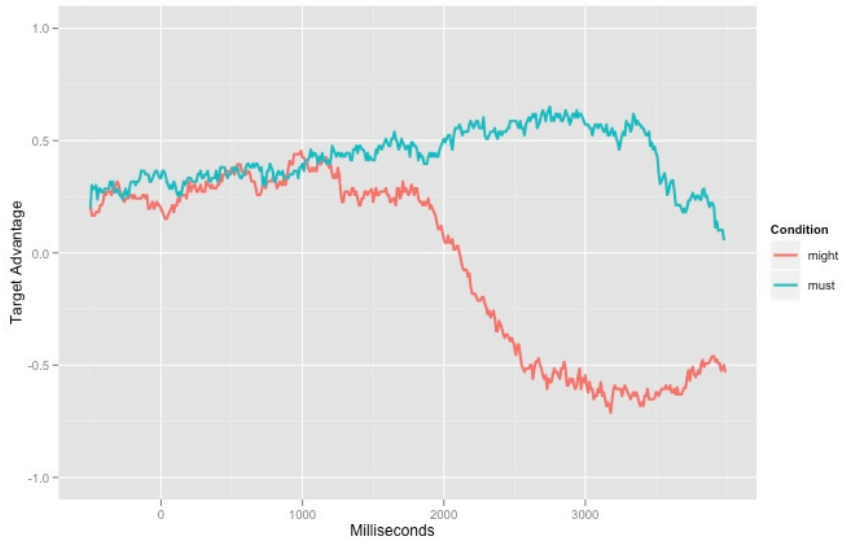

Figure 4: Must-Might experiment: Target advantage (must- might) starting from 500ms prior to modal onset until 2 seconds after column onset: around 1000ms after the modal onset there is a clear divergence between the *must* and the *might* conditions.

This figure shows that starting from 1000ms after modal onset the *must* and *might* conditions diverge, as the target advantage score for the *might* condition lowers. Before that the target advantage score for both conditions is (equally) high, meaning that participants looked more at the shape that could be guessed with certainty for both conditions. Thus, subjects initially looked at the target (the shape that could be guessed with certainty) more than they looked at the competitor (the shape that could not be guessed with certainty). Upon hearing "must" subjects kept looking at the target, and 1000ms after hearing "might" subjects started looking more at the competitor. If we take the display in figure 3 as an example, this would mean that subjects initially looked more at the target (the shape in the upper right, which could be guessed with certainty). When the confederate uttered *There must be a red square located in the upper right*, upon hearing "must", participants kept looking more at the target; when the confederate uttered *There might be a red square located in the bottom left*, 1000ms after the onset of "might" participants started looking more at the competitor (the shape in the bottom left, which could not be guessed with certainty).

In the *might* condition subjects began directing their looks towards the appropriate shape only about 1000ms after the modal onset, and prior to that, they were looking at the shape that could be guessed with certainty. Taking into account that about 200ms are necessary to plan and preform a saccade, this can be interpreted as an 800ms delay in looks to the correct shape for the *might* condition.

For items with correct *might* guesses the total ambiguity period was divided into two windows: one from the modal onset and until 1 second after the modal onset (window A), and the other from 1 second after the modal onset and until 200 milliseconds after the onset of the row

in their privileged ground. On average, the shape onset started 869ms after the modal onset. Since 1000ms is when the effect of the modal type begins to become evident, having an incorrect shape type guess could easily mask such an effect, explaining why it would be present only when the guess is correct.

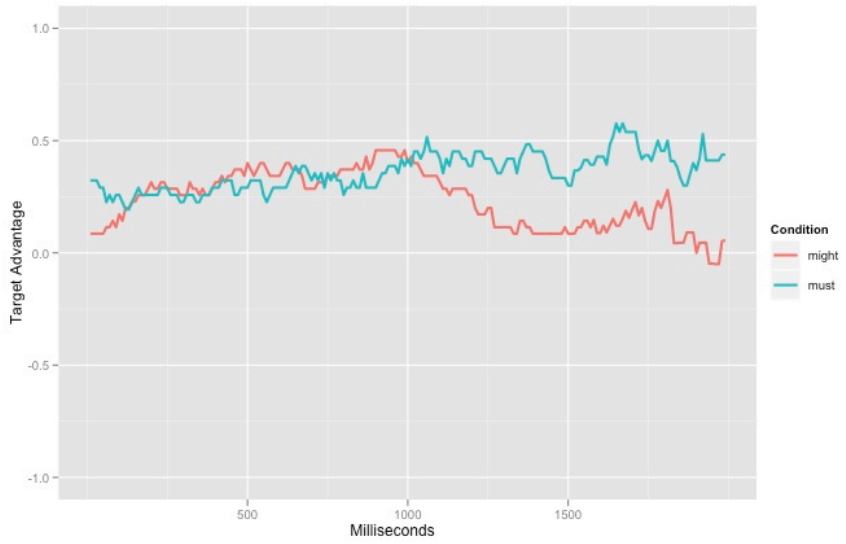

Figure 5: Must-Might experiment: Target advantage (*must- might*) starting at the modal onset and until 200 milliseconds after the onset of 'row' for items with correct guess.

locator (window B). ANOVAs were done on the target advantage scores using these windows above on subject and item means. There was a significant interaction between time window and condition, both by subject ($p<0.05$) and by item ($p<0.01$).

Discussion About 1 second after the modal onset, target advantage scores for the *might* and *must* conditions diverged. Prior to that, the score for the *might* items patterned with that for the *must*. This result can be interpreted as evidence that implicatures processing is slow: prior to processing the implicature, *must* patterned with *might* and diverged from it when the implicature was processed. Taking into account 200ms for planning and performing a saccade, this results in 800ms for processing the implicature in *might*.

When interpreting the results, however, one has to take into account that there is an initial target bias. Throughout the entire ambiguity period the target advantage for both conditions is above zero. This is true for both when the guess is correct and when it is not (for the *might* condition). There seems to be an initial preference to look at the shape that could be guessed with certainty even for the *might* condition, even when there was no independent reason to prefer the target. This preference was larger when must was uttered, but it was nevertheless present for the *might* condition as well. Furthermore, in the first second after modal onset, there was no significant difference between *might* and *must*, and for both there was a preference to look at the target item. A possible explanation is that since the shape that could be guessed with certainty was easier to guess, the subject might have anticipated that this shape would be guessed first, and hence this shape attracted more initial looks. The items for which the *might* guess was wrong add further to the targets bias. The target advantage score for the *might* conditions in which the guess was correct drops around 1000ms after the modal onset. For items in which the guess was wrong the drop in target advantage for the *might* condition is slower, and patterns with that for the *must* condition until about 2 seconds after the modal onset. As mentioned above, when the guess for the *might* condition was wrong, the shape that was guessed matched the shape that was placed in the target position, and hence this resulted in increased looks to the target for the *might* condition

in items in which the guess was wrong. This initial target bias makes it more difficult to show that there was no delay in the processing of "must": if there is an independent initial preference to look at the target item, it would be difficult to tell when looks to the target item are prompted by hearing must and when they just follow from an independent preference. Independent evidence that *must* is processed immediately is therefore needed[11].

Additional issue with the above results is the difficulty of the task. In order to attribute the effect found to the implicature involved in *might*, other factors (such as difficulty of the task) need to be ruled out. The task subjects needed to perform was fairly complicated: they needed to be able to do the kind of reasoning (using the game rules) that would allow them to realize that one shape could be guessed with certainty while the other shape could not. To aid them with this task before the target trials took place there was a "training" period, in which the subject was given similar displays and had to use the game rules to make guesses (there were 16 such trials for each subject), and while I believe that this indeed helped subjects prepare for the task in target trials, the extent to which people seemed to absorb the rules varied across subjects (some people did better than others when playing the "guesser" themselves). Taking this into account, since there was an effect of the condition (whether the confederate said *must* or *might*) on looks, it is reasonable to assume that people learned the game rules to the extent to which the experimental task yielded meaningful results.

However, I think that it will still be useful to replicate the above results with an easier task, and follow up with a study that directly compares modals to quantifiers and adjectives.

4.2 Follow-Ups

I propose two additional studies: a study that asks the same question as the *must-might* study but uses a more simple task, and a study that examines the time course of implicatures triggered by gradable adjectives.

4.2.1 Replication of the *must-might* study with a simpler task

This study is designed to examine again the time course of the implicature involved in the processing of *might*, but with a simpler task. Subjects will be presented with an array of three objects: two small ones and one big object (figure 6). Then they will be told that a fictional character is going to arrange them in boxes, one item in each box. At this point the display will change and they will see a display of three containers: two of the containers will be small and could only contain one of the small objects, but one container will be big enough to contain the big object. Thus the location of the big object could be inferred with certainty (it is in the big container) but the location of the other two smaller objects cannot be predicted- since they can be in either one of the small containers.

Then subjects will be prompted to answer one of the following questions:

1. Where *might* the [small object] be?
2. Where *must* the [big object] be?

[11] In a pilot study comparing *must* to *is* indeed there was no delay associated with *must*. However, the results were not replicated in a complete follow up study.

Since only one container is big enough to contain the big object, upon hearing *must* subject should realize that they are being asked about the one object whose location they can tell with certainty and start looking at the big container. However, since they would not be able to tell with certainty where each of the small objects is, *might* should trigger looks to the smaller containers and away from the big container.

An example of a display for a target item is presented in figure 6.

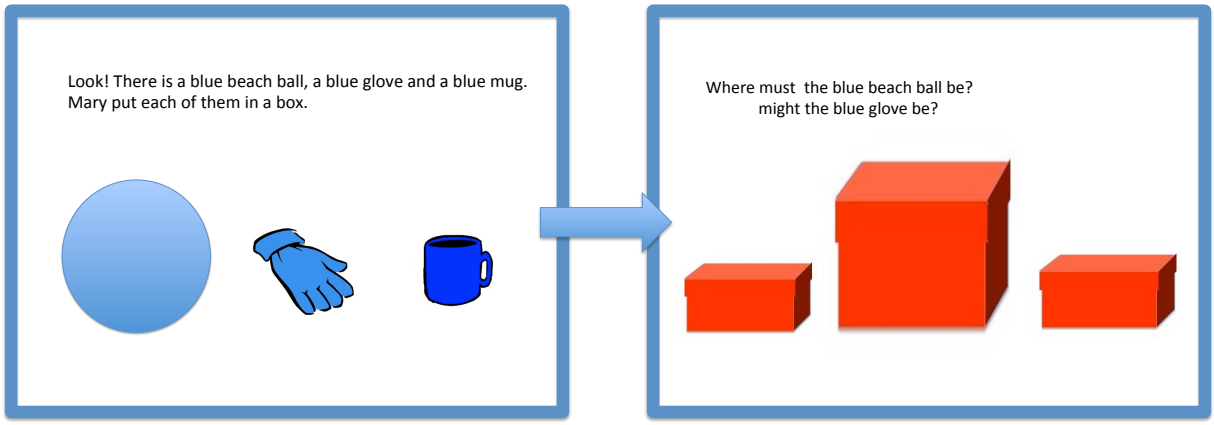

Figure 6: Example of a target trial in the new must-might study.

The second display will be accompanied with the following (pre-recorded) utterances:

1. Where *must* the blue beach-ball be?
2. Where *might* the blue glove be?

Before the onset of *beach-ball* or *glove* nothing but the usage of the modal indicates which one of the containers the subject is being asked about[12].

If indeed *might* involves a processing delay induced by computing an implicature, whereas *must* is interpreted immediately the following results are expected:

- *Must* will trigger immediate looks to the big container.
- When *might* is uttered people will at first look at random at the three containers, and only after some time they will start looking at the two small containers and away from the big container.

Since this task is much simpler than the task in the above reported study, replication of the reported results will provide strong evidence towards delayed implicature processing.

[12]This is also the reason that the utterances are phrased as questions. Since in the first condition there is only one box in which the big item can be, a paraphrase as a non-question would be *"Click on **the** box that must contain the blue beach ball"*. In the second condition, since there are two possible containers the instructions will have to be phrased as *"Click on **a** box that might contain the blue glove."* Hence early on (before hearing the modal) the subjects would have a disambiguating clue (whether "a" or "the" was uttered). This issue was resolved by phrasing the instructions as questions.

4.2.2 Gradable Adjectives

Doing a study about implicatures triggered by gradable adjectives may turn out to be more difficult than expected. This is simply because there are few cases of adjectives that have lexicalized intermediate values (*hot*, *likewarm* and *cold* is one such example). Furthermore, in many cases intermediate values may be associated with attempts of being polite rather than actual denotation (for instance, people will often say about someone that they are *chubby* instead of *fat* in order to avoid hurting that person's feelings). Hence examples directly isomorphic to the none-some-all scale might be difficult to find. However, if instead of insisting on using scales with lexicalised intermediate values we admit using *somewhat x* for the intermediate value, we might be able to get around the above specified issue. For example, *somewhat tall* has the same type of reasoning associated with it that *some* has. If Joe is somewhat tall, then his height is less than the height of someone I would call *tall*.

Therefore I propose a study that will examine the time-course of implicatures triggered by gradable adjective by using stimuli of the following type. While viewing the scene in figure 7 subjects will hear the following instructions:

1. Click on the mug that is tall.
2. Click on the mug that is short.
3. Click on the mug that is somewhat tall.

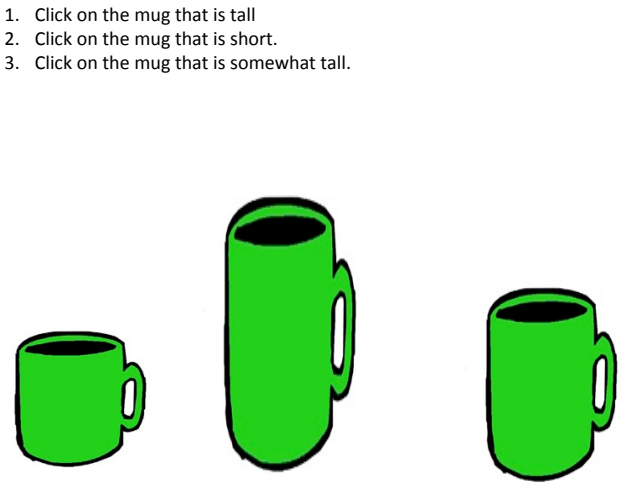

Figure 7: Example of a target trial in the adjectives study.

If *somewhat tall* involves the same type of implicature that *some* has (i.e., not quite 'tall enough') then in the last condition looks to the target (the intermediate mug) will be slower than in the first two conditions.

The time-course of the adjectival implicatures can then be compared to the time-course of the modal-implicatures from the above study. If the results are similar then this would serve as

evidence for treating modals like adjectives. If the results vary in a significant manner then this would serve as some evidence against treating modals like adjectives.

5 Reasoning about *must*

There has been much debate in the literature about whether *must* is weaker than *is* in terms of the amount of certainty that it conveys (Veltman [1985]; Kratzer [1991]). For instance, if I say *"It must be raining."* then people would have the feeling that I am somewhat 'less certain' than if I said *"It is raining"*. Under Lassiter's scalar (or threshold) account, *must* is in fact 'weaker', since one can utter *must p* if the probability of p is higher than the threshold $1 - \alpha$ where α is contextually given.

In a recent paper von Fintel and Gillies [2010] provide an analysis of *must* treating it as an evidence marker. The bulk of their analysis is that the difference between *must* and *is* is really rooted at the fact that *must* signals that the speaker is making an inference, and hence the information provided is based on indirect (rather than direct) evidence. In that sense, *must* is not really about making a weaker statement, but it is rather about signaling that the information provided was attained by making an inference. In that sense, there is nothing weaker about *must*, and it "stays strong". Hence a corollary of that account would be that if we were to arrange modal expressions (and *is*) on a scale *must* would be associated with the same threshold value that *is* would have been associated with (whether we would like to say that this is 1 or a little bit lower than 1, at any case, higher than Lassiter's $1 - \alpha$).

The results from the Honest Joe study reported earlier however point to a *must* being 'weaker' than *is*. People consistently placed smaller bets when the information was presented with *must* than when it was presented with *is* (and there was a significant difference between the two conditions), suggesting that statements presented with *must* are taken as weaker than those presented with *is*. On a first glimpse this seems to contradict the analysis proposed in von Fintel and Gillies [2010].

There is a way to reconcile the results from the Honest Joe experiment with the analysis of *must* proposed in von Fintel and Gillies [2010]. The main point is that *must* does indeed involve making an inference, and that inference is what makes it weaker. This is because unlike inferences made in a logical system (proofs) inferences made in an everyday conversation are less reliable.

Inferences made in first or second order logic are proofs. Hence if I can derive via a proof that q_n follows from a set of given premises $\{q_1, ..., q_{n-1}\}$ then there is nothing weaker about q_n, it is not is any less certain than either one of $q_i, 1 \leqslant i \leqslant n-1$. However, inferences made by people are generally not proofs in a logical system. People often make faulty conclusions, either based on various logical errors, on incorrect assumptions, etc. We do not even need to go as far as errors in reasoning, since people often reason "probabilistically" in the sense that they are ready to draw conclusions (and use *must*) even when the information they are basing themselves on is less than certain (7), or when the entailment relation is less than certain (8).

(7) Mary is usually in her office on Mondays, and today is Monday, so she **must be** in her office.
*Mary is **usually**, but not always in her office on Mondays. She might not be there now, because she might have taken a vacation or she might be sick. She might even be out for lunch.*

(8) I can't remember where I put my keys. They're not in my purse and neither in my pocket. I **must have** left them in the car then.
The three most likely places for my keys are my purse, my pocket or the car. But that does not mean that the keys really cannot be anywhere else except in one of these places. For all I know they might have fallen out, or somebody might have taken them. The inference I make, namely that if the keys are not in my pocket or in my purse then they are in my car is probabilistic. In reality, if the keys are not in my pocket or in my purse then they are most likely to be in the car, but they are not ***necessarily*** *there.*

Hence the status of conclusions made by making inferences in everyday life is in fact weaker than that of information attained directly (i.e., via one of the senses). If *must* signals that an inference is being made, then it does in fact signal that the information offered is less reliable.

There are then two theoretical parts to be examined. The first one is whether in fact *must* signals making an inference. The second part is whether statements made with *must* are considered by the hearer as less reliable than statements made without it (using *is* for example).

In a a series of eye tracking studies (a pilot and an attempt to replicate the pilot study) I examined the first part by manipulating common vs. privileged ground. The setting was the same as in the *must* and *might* experiment reported above: a 'guessing game' in which one of the participants (the verifier) had complete information and the other participant (the guesser) had only partial information about a visual scene consisting of colored shapes. Hence part of the information was privileged (only available to the verifier) and the rest was common ground (known to both participants). If *must* is indeed associated with making inferences, then a verifier would assume that if the guesser uses *must* (rather than *is*) it is because he does not have direct evidence and is hence referring to the information that is privileged for the verifier.

In this study too there were two occluded shapes in critical trials. This experiment however included one further manipulation, as the guesser was allowed to ask to be told what one of the hidden shapes is (the verifier would then answer the question, but without revealing the shape on the screen).

Target items consisted of 12 sets of scripted utterances, corresponding to 12 scenes. The confederate read off his screen the scripted utterances while pretending that these were spontaneous guesses. The utterances referred to the two hidden shapes in each scene. The confederate asked about one of the hidden shapes, and then made a guess regarding the other shape. For example, the utterances corresponding to the scene shown in figure 8 are:

What shape is in the bottom right? (a blue heart)

1. There *must be* a blue heart located in the bottom left, because you said there was a blue heart in the bottom right.

2. There*'s* a blue heart located in the bottom right, so I think there is a blue heart in the bottom left.

A "must (be)" always referred to the shape the confederate had to guess, while "is" always referred to the shape about which the confederate was already informed. Whether the confederate was referring to the shape that he was about to guess or to the shape that was already orally revealed became evident when the shapes location was revealed. The phrase

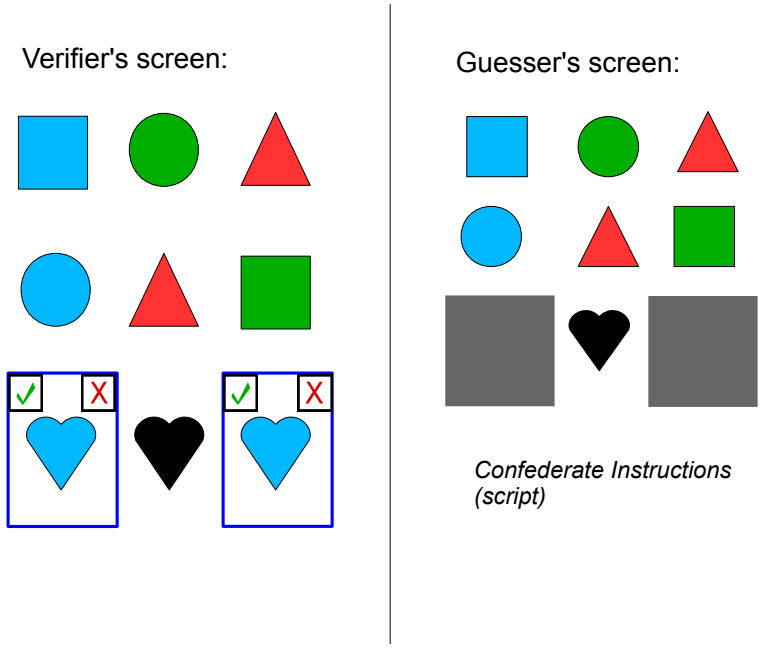

Figure 8: Must-Is experiment: Example of a target trial.

'located in' was added to extend the uncertainty period. The conditions were such, that once the confederate had asked about one of the hidden shapes in the target trials, he could then use this information with combination of the rules to guess what the other hidden shape was with certainty. The subjects eye gaze was recorded, in order to compare looks to the shape whose content was already orally revealed vs. the shape that he confederate had to guess. The prediction was that when "must" was used subjects would look more to the shape that was not yet revealed, and was still in the subjects privileged ground; hearing "is" in contrast would trigger more looks to the shape that had already been revealed, and was in the common ground.

Below I report the results from a pilot study and an attempt to replicate it. *Results: Pilot study.* 6 undergraduates participated in the pilot study in exchange for course credit. The target was taken to be the shape the content of which was not yet revealed (corresponding to the must), and the competitor the shape that had its content orally revealed (corresponding to the is). There are two conditions: must, corresponding to when the subject heard must, and is, corresponding to when the subject heard is. Target advantage score was computed as looks to the target minus looks to the competitor. This measure is appropriate since it shows more clearly the preference of the target over the competitor.

Figure 9 shows the target advantage from 500ms before the onset of the modal and until 2 seconds after the onset of the column word, for the two conditions: *must* and *is*.

This figure shows that starting at around 200ms after the modal/is onset there is a clear divergence between the must and the is conditions. The ambiguity period was defined to be starting from the onset of must or is, and until 200ms after the onset of the row locator. Figure 10 shows the target advantage for this period.

These results show a clear divergence between the *must* and *is* condition, at about 200ms

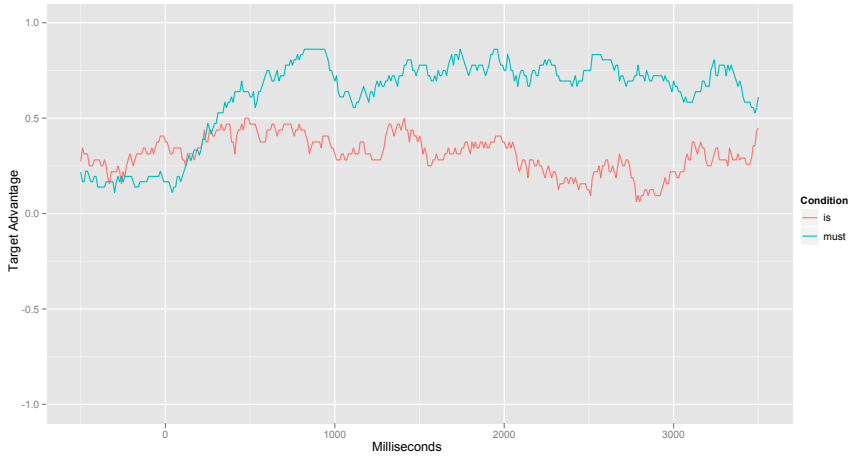

Figure 9: Must-Is experiment: target advantage from 500ms before the onset of the modal and until 2 seconds after the onset of the column word.

after modal onset hence providing evidence for immediate processing of *must*,

Results: complete study. In a full study I attempted to replicate the above results. The study however did not replicate. I collected data from 11 subjects, who participated in the study in exchange for extra credit.

Target advantage score 500ms prior modal/is onset and until 2s after onset of column is shown in figure 11

In the first second after modal onset looks to the target were higher for the *is* than for the *must* condition.

It seems that the main problem was that in the context was one of guessing. In a context of guessing even a utterance with *is* was taken to denote a guess. It is still surprising however that *is* would trigger more looks to the target than *must* did. It is also worth noting that even though target advantage was higher for *is*, it was above 0 for the *must* condition as well (meaning that nevertheless subject looked more at the target than at the competitor upon hearing *must*). What still remains difficult to explain is why this did not seem to be an issue with the pilot study.

Discussion The pilot study suggests that people are sensitive to whether *must* or *is* was used, and looked more at that part of the screen that required making an inference than the part that was already discussed upon hearing "must". These results however did not replicate in the complete study. The problem seemed to be that in the context of making guesses *is* too was interpreted as a guess. Thus it seems to me that a better way to test this is to have a context in which both 'guesses' and 'non guesses' are equally expected.

6 Number sense and modals

There is a growing body of work on how people interpret quantifiers (Petroski et al. [2009]; Clark and Grossman [2007]; Troiani et al. [2007]), and to what extent this capability is related to peoples' number sense. The idea in the quantifier-reasoning experiments is to learn

Figure 10: Must-Is experiment: Target advantage for the ambiguity period.

something about quantificational reasoning by asking subjects to verify a statement featuring a quantifier given visual stimuli with varying complexity (i.e., varying number of items) which was shown for short periods of time.

Work on quantifier reasoning shows connection between number sense and the interpretation of quantifiers. For instance, Clark and Grossman [2007] show that impaired number sense can result in impaired ability to interpret sentences involving quantifiers. Petroski et al. [2009] offer some evidence that *most* is understood in terms of cardinality comparison (and this was so even when counting was made impossible because the visual stimuli was only presented for a very short time).

If modals are quantifiers then we should be able to find a similar relation to the number system. For instance, impaired number sense should also cause some impairment in the ability to interpret modal statements. And we should also be able to find some evidence that when people evaluate modal statements reasoning about cardinalities is involved in a similar way.

I propose to use stimuli and methods similar to the ones in the study in Petroski et al. [2009] to gain better understanding of how expressions of modality are interpreted, and if and how they are related to the number sense. This can then serve as some evidence for (or against) treating modals as quantifiers.

The design is inspired by the design of the study about 'most' in Petroski et al. [2009]. Subjects will see a display of 16 circles on the screen each time, some of which will be black and some white. The number of white and black circles will vary. They will be then instructed that either the computer will select a circle at random and they will have to answer a question regarding the computer's selection, or they will just have to answer a question about the display.

There are two settings: a quantifier and a modal setting. In the quantifier setting subjects will read a sentence: *All/some/more than half of the circles are white.* Then the display will change and they will see a scene consisting of 16 circles. Each of the quantifiers will appear with each one of the four display types:

Figure 11: Must-Is replication experiment: Target advantage for the entire period.

1. All circles are white (A)
2. Exactly half (8) of the circles are white (B)
3. 10 out of the 16 circles are white (C)
4. All but two of the circles are white (D)

An example of the four display types is shown in figure 12.

In the modal setting the displays will be the same, except *all* will be replaced by *must*, *some* will be replaced by *might* and most will be replaced with *should*.

This results in 24 conditions {display type} x{linguistic expression (*all, more than half, some, might, should, must*). Instead of limiting the amount of time the stimuli is presented (like in the Petroski et al. [2009] study, in which scenes were shown for 200ms.), I propose to let subjects view each display as long as they want to, but measure their response times (and ask them to respond as fast as they can). Then the response time can be looked at as an additional means of comparing quantifiers to modals.

I also propose to have some fillers which will consist of similar displays (some of the circles will be white and some black), but instead of *some/all/more than half* the sentence will use an exact number for half of the fillers. For the other half the modal statement will be replaced by *There is an X% chance that the computer will select a white ball.*

Response type ("Yes/No") and response times will be recorded. If modals are inherently quantifiers I expect *must* to parallel *all* and *might* to parallel *some*. That does not necessarily mean that *must* should behave exactly like *all* and that *might* should behave exactly like *some*. However, I expect there to be similar general effects.

For instance, if response times are slower for *"Some of the circles are white"* while viewing scene A (due to a violation of the *not all* implicature), and rejection rates are higher, then

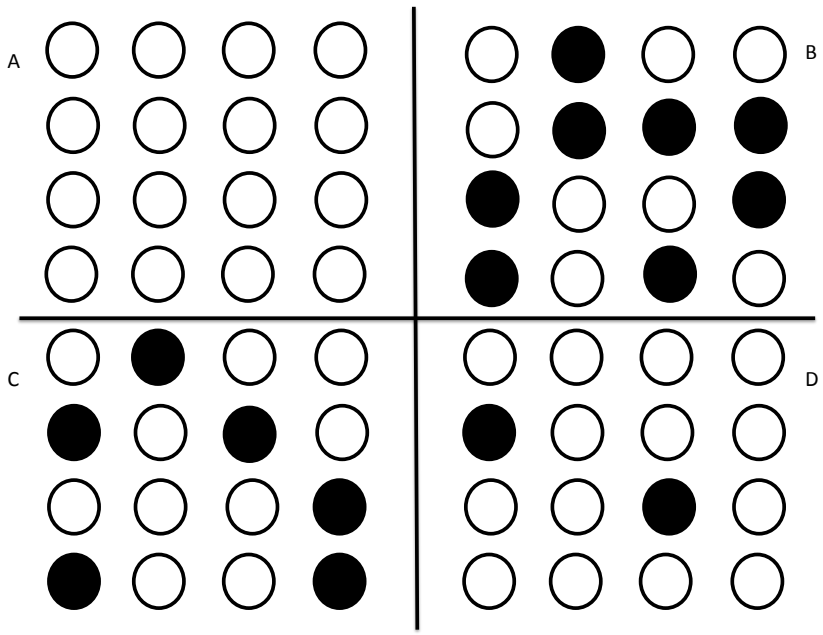

Figure 12: The four displays: different number of circles are white

I expect to find a similar effect when the scalar statement is replaced with *"The computer might select a white circle".*

It is even more interesting to see what happens when a modal with an 'intermediate' strength, such as *should* is used, and compare it to the *more than half* quantificational case. I expect statements with *more than half* to be the hardest ones to evaluate in scenes of type B and C, and thus to be associated with the biggest response time among the quantifiers. It would be interesting to see whether using modals with intermediate strength (such as *should*) would have a similar effect.

7 Summary and timeline

7.1 Summary

In this work I address various aspects about reasoning with modals. In particular, current and proposed research address the issue of scalar implicature processing when the trigger is a modal expression, as well as comparing that to quantifier and gradable adjective implicature triggers. These experimental results then can help us distinguish between different theoretical issues (i.e., whether it makes more sense to think about modals as quantifiers, or as gradable scalar expressions). I have also presented empirical results motivating ranking between different modal expressions, and proposed ways to extend that research to try and answer questions about varying amounts of certainty and perspective taking (i.e., *certainty gaps*). Finally, the experimental results from "Honest Joe" and the "must-is" experiments can shed some light on the difference between *must* and *is* in terms of inference making, and how inference making relates to judgements of certainty.

7.2 Timeline

By the end of August 2012 I intend to have followed up on any concerns brought up by the committee, to have read any additional suggested literature and to have completed and implemented the experiments described in this proposal, and any additional experiments that may seem appropriate. By the end of December 2012 I intend to have collected and analyzed all necessary data. The remaining part of the next academic year (starting from January 2013) will be dedicated to the write-up: by January 2013 I plan to be regularly submitting chapters to members of my committee, and I aim for having a complete dissertation by the end of summer 2013.

References

J. Baron. *Thinking and deciding*. New York: Cambridge University Press, 2004.

J. Baron and J. C. Hershey. Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, 54:569–579, 1988.

M. H. Birnbaum and A. Chavez. Tests of theories of decision making: Violations of branch independence and distribution independence. *Organizational Behavior and Human Decision Process*, 71:161–194, 1997.

S. Brown-Schmidt, C. Gunlogson, and M. K. Tanenhaus. Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, 107:1122–1134, 2008.

R. Clark and M. Grossman. Number sense and quantifier interpretation. *Topoi*, 26:51–62, 2007.

P. Grice. Logic and Conversation. *Syntax and Semantics*, 3:41–58, 1975.

J. Hintikka. Semantics for propositional attitudes. *Models for Modalities*, pages 87–111, 1969.

L. Horn. Toward a new taxonomy for pragmatic inference: Q-and r- based implicature. *Meaning, Form and Use in Context*, pages 11–42, 1984.

L. Horn. Implicature. *The Handbook of Pragmatics*, pages 3–28, 2004.

C. Kennedy. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45, 2007.

B. Keysar, D. J. Barr, J. A. Balin, and J. S. Brauner. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38, 2000.

A. Kratzer. What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, 1(3): 337–355, 1977.

A. Kratzer. The notional category of modality. *Words, worlds, and contexts: New approaches in word semantics*, pages 38–74, 1981.

A. Kratzer. Modality. *Semantics: An international handbook of contemporary research*, pages 639–650, 1991.

A. Kratzer. Modality and conditionals: New and revised perspectives. To Appear., 2012.

S. Kripke. Semantical considerations on modal logic. *Acta philosophica fennica*, 16:83–94, 1963.

M. Larson, R. Doran, Y. McNabb, R. Baker, M. Berends, A. Djalali, and G. Ward. Distinguishing the said from the implicated using a novel experimental paradigm. *Proceedings for Experimental Pragmatics*, 2007.

D. Lassiter. *Measurement and Modality: The Scalar Basis of Modal Semantics*. PhD thesis, Department of Linguistics, New York University, 2011.

L. Lopes. When time is of the essence: Averaging, aspiration, and the short run. *Organizational Behavior and Human Decision Process*, 65:179–189, 1996.

P. Petroski, J. Lidz, , T. Hunter, and J. Halberda. The meaning of 'most': Semantics, numerosiy and psychology. *Mind & Language*, 24(5):554–585, 2009.

V. Troiani, J. E. Peelle, C. Halpern, R. Clark, and M. Grossman. Dissociable numerosity and executive components of quantifier knowledge. *Brain and Language*, 103, 2007.

A. Tversky and D. Kahneman. Advances in prospect theory: cumulative representations of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992.

F. Veltman. *Logics for conditionals*. PhD thesis, University of Amsterdam, 1985.

K. von Fintel and A. S. Gillies. 'might' made right. 2007.

K. von Fintel and A. S. Gillies. Must... Stay... Strong! to appear in *Natural Language Semantics*, 2010.

K. von Fintel and S. Iatridou. How to say ought in foreign: the composition of weak necessity modals. Paper presented at the 2006 Workshop on Philosophy and Linguistics, Ann Arbor, University of Michigan, 2006.

S. Yalcin. Probability operators. *Philosophy Compass*, 5(11):916–973, 2010.