An experimental investigation of the weakness and evidentiality of epistemic *must*

Languages allow for a wide variety of ways to communicate both about a) one's beliefs as well as about b) one's evidence for those beliefs. While languages like Turkish (XXX citation) and Quechua (Faller XXX) have rich evidential systems by which a speaker marks an assertion for whether the evidence supporting the assertion was obtained, e.g., directly, via hearsay, or through inferential means. In contrast to these rich evidential systems, the English evidential system is relatively impoverished, or rather, the lexical items used to mark evidentiality in English usually have other uses as well, which makes it difficult to tease apart a given lexical item's evidential from its assertoric contribution. Here we focus on one such lexical item, the case of epistemic *must*.

Since Karttunen (1972), linguists have debated the meaning of this word. The one clear empirical fact seems to be that an utterance of *must q*, as in (2), does not entail that *q*, as in (1).

(1) a. It's raining.

b. It must be raining.

This is surprising under a strong necessity semantics for *must*, which predicts that such an entailment relation *should* exist. Since (XXX citation), different accounts of the weakness of *must* have been put forward. Some have proposed that the weakness is in the semantics of *must* itself (Kratzer, Lassiter, XXX), while others have proposed that the semantics of *must* is strong, but its weak interpretation is the result of an inference about the evidential status of p (vFG XXX). In particular, the claim has bene that *must q* communicates that the speaker's evidence for *q* is indirect or inferential (as defined in XXX citation's tree), or that the evidence is below a certain strength threshold (Matthewson, XXX). Others have tried to frame the strength of *must* as an issue of speaker commitment: a speaker who utters *must q* is committed to the truth of *q*, given indirect evidence for *q* (vFG, XXX), though citing naturalistic corpus examples, Lassiter (2015) observes that speakers need not be fully committed to *q*; the strength of their belief in *q* must simply be greater (by some large margin) than the belief in any alternative, given the available indirect evidence.

A big problem for the treatment of *must* is that many of the different ways of treating its weakness and evidentiality cannot be teased apart empirically. Here we make headway on the parts that can, as well as proposing an alternative formal, implemented, model of *must* that derives its weakness as an M-implicature: *must q* is marked (i.e., costly) relative to the bare form (1a); since the bare form is sufficiently strong to communicate *q*, listeners weaken the interpretation of *must q*. This account is implemented within the Rational Speech Act framework ((?, ?, ?)), which has the advantage of being explicit about representing speaker and listener *beliefs*, and in turn making predictions for both *production* and *interpretation* choices.

Empirically, we address the following questions: (i) Is the use of *must q* only felicitous with indirect evidence, or with evidence whose strength is below a certain threshold, or is the probability of using*must q* probabilistically modulated by evidence strength (Exp. 2)? (ii) Does *must q* result in weak listener belief in *q* compared to bare *q* (Exp. 3a)? (iii) Does *must q* commit the speaker to *q* (Exp. 3b)? The M-implicature model, which we present below, predicts XXX especially continuous evidence strength

In **Exp. 1 (n=40)**, we collected estimates of evidence strength. Participants on Amazon's Mechanical Turk rated the probability of *q* (e.g., of rain) given a piece of evidence *e* (e.g., *You hear the sound of water dripping on the roof*) on a sliding scale with endpoints labeled "impossible" and "certain". These estimates were used for analysis in Exps. 2 and 3.

Exp. 2 (n=40) tested how likely speakers are to use the marked *must q* utterance as evidence strength decreases. On each trial, participants were presented with a piece of evidence (e.g., *You see a person come in from outside with wet hair and wet clothes*) and were asked to choose one of

four utterances—bare (1a), *must q* (1b), *probably p*, *might p*—to describe the situation to a friend. Participants were more likely to choose the more marked *must* form over the bare form as the strength of evidence decreased ($\beta = 5.4$, $SE = 2.4$, $p < .05$), even when controlling for evidence type (e.g., perceptual, reportative, inferential). Importantly, there were cases of direct perceptual evidence for q in which participants nevertheless chose the bare q utterance.

Exp. 3a (n=120) tested whether listeners' estimates of a) the probability of q and b) the strength of speakers' evidence for q differ depending on the observed utterance; i.e. whether listeners take into account their knowledge of speakers' likely utterances in different evidential states as they interpret the bare and *must* forms. On each trial, participants were presented with an utterance (e.g. *It's raining*), and were asked a) to rate the probability of q on a sliding scale with endpoints labeled "impossible" and "certain"; and b) to select one out of five pieces of evidence that the speaker must have had about q. Participants' believed q was less likely after observing the *must* utterance ($\mu = .65, sd = .21$) than after observing the bare utterance ($\mu = .86, sd = .15, \beta = -.21$, $SE = .02$, $t = -10.1$, $p < .0001$). In addition, average strength of evidence was lower after *must* ($\mu = .78, sd = .12$) than after the bare utterance ($\mu = .87, sd = .1, \beta = -.08$, $SE = .01$, $t = -6.8$, $p < .0001$).

Exp. 3b (n=60) tested listeners' judgments of the speaker's commitment to q. The procedure was the same as in Exp. 3a, with a minor variation in the dependent measure: participants were asked to rate the probability of the speaker believing q on a sliding scale with endpoints labeled "impossible" and "certain". Participants' rated speakers as more strongly believing in q after observing the bare utterance ($\mu = .96, sd = .07$) than after observing *must q* ($\mu = .78, sd = .18, \beta = -.2$, $SE = .02$, $t = -12.42$, $p < .0001$). A comparison of the data from Exps. 3a and 3b revealed that participants judged speaker commitment to be stronger than their own resulting belief in q ($\beta = .11$, $SE = .02$, $t = 5.39$, $p < .0001$). Nevertheless, speaker commitment for *must p* was not judged at ceiling, supporting ? (?) but not ? (?).

Taken together, these results support an M-implicature account of the choice and interpretation of epistemic *must*: the longer, marked, *must* is interpreted by listeners as conveying the marked meaning that the speaker arrived at the conclusion that q via an evidentially less certain route than if they had chosen the shorter, unmarked, bare form.

Following ? (?), we present an extension of the Bayesian Rational Speech Act framework using lexical uncertainty (cite) to derive the implicature. In this model, the semantics of the bare utterance and *must q* are relatively unconstrained. We define the semantics of the utterances such that $p(q|bare) > \theta_b$ and $p(q|must) > \theta_m$, where the pragmatic listener is uncertain about θ_b and θ_m and infers the values through pragmatic reasoning. When the cost of uttering *must q* is greater than the bare form, the pragmatic listener infers that $p(q)$ is smaller than when the utterance is the less costly *bare q*. Given the weakened certainty of q, the listener may then infer that the speaker has weak or imperfect evidence of q. Our empirical results and computational model support this account and provide a new perspective on the meaning of *must*: its weakened meaning derives from straightforwardly from an M-implicature.

say sth reasonable about how this sort of model lets us move from qualitative predictions of unimplemented formal models to quantitative predictions of implemented formal models. also say sth about which theories from intro this stuff bears on (e.g. in support of dan's claim that it's not full speaker commitment (against vFG)); but also: against any theory that makes threshold predictions for evidence or is uni-dimensionally focused on evidence type, though there's evidence (ha!) that evidence type matters, too, above and beyond evidence strength. maybe also say sth critical about how to proceed, in the spirit of the workshop?

References

Karttunen, L. (1972). *Possible* and *must*. In J. Kimball (Ed.), *Syntax and Semantics*, Volume 1, pp. 1–20. New York: Academic Press.