The weakness of epistemic *must*: A pragmatic reasoning approach

Judith Degen (jdegen@stanford.edu), Gregory Scontras (scontras@stanford.edu), Andreas Trotzke (andreas.trotzke@uni-konstanz.de), Eva Wittenberg (ewittenberg@ucsd.e

October 13, 2015

INSERT

Abstract

Keywords: pragmatics; semantics; psycholinguistics; modals; discourse particles; German; English

1 Introduction

XXX

In the following, we label English and German experiments as *E.N* and *G.N*, respectively.

2 Experiment E.1: evidence strength

In Exp. E.1, we collected estimates of evidence strength.[1] These estimates were used in the analysis of Experiments 2 and 3.

2.1 Methods

2.1.1 Participants

40 participants were recruited through Amazon's Mechanical Turk crowd-sourcing service, and were compensated for their participation.

2.1.2 Materials and procedure

Participants rated the probability of a state of affairs q given a piece of evidence e by adjusting a slider on a scale with endpoints labeled "impossible" and "absolutely certain". On each trial, participants first saw the context sentence "Imagine that you are at home." Then the evidence e for q was shown, e.g., "Dinner is usually ready around 6pm. You look at the clock and it is 6pm." Finally, participants were asked about the probability of q, e.g., "How likely is it that dinner is ready?" and adjusted the slider accordingly. There were four different q that appeared in the "How likely is it that q?" frame:

(1) it is raining

[1] This experiment can be viewed here.

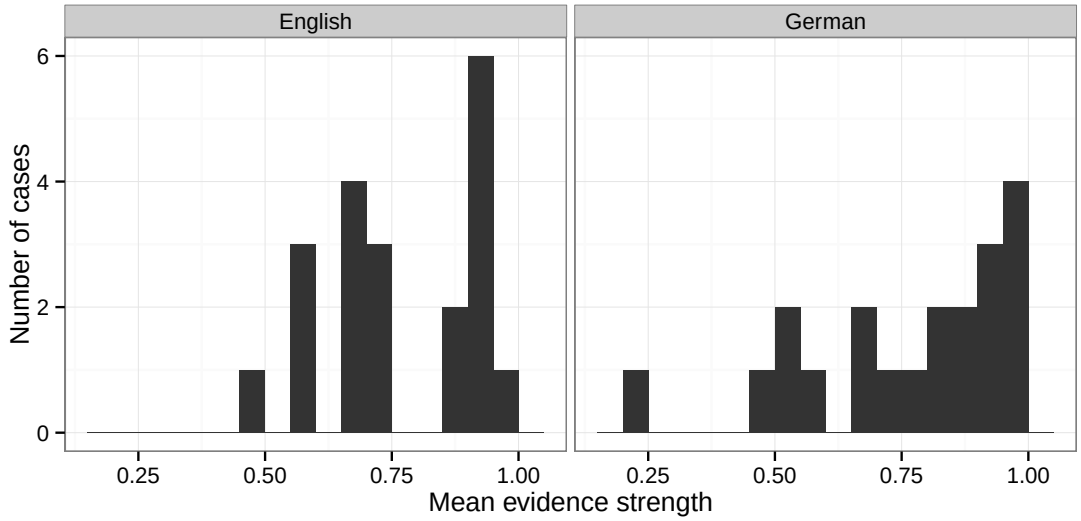

Figure 1: Histogram of by-item evidence strength means for English (Exp. E.1, left) and German (Exp. G.1, right).

(2) the coffee is cold

(3) dinner is ready

(4) the neighbor's dog is barking

For each q, each participant evaluated one of five possible pieces of evidence, resulting in four trials per participant. Trial order was randomized.

Pieces of evidence were generated by taking a version of the top five most frequently generated explanations from a separate experiment in which q was given to participants and they were asked to provide a free response explanation of how the speaker knows about q.[2] The full list of pieces of evidence for each q can be found in Appendix A.

2.2 Results and discussion

We obtained between 3 and 14 ratings for each piece of evidence. We interpret the slider value between 0 ("impossible") and 1 ("absolutely certain") as a participant's estimate of the probability of q, given e, which we will also sometimes refer to as *evidence strength*. A histogram of mean evidence strengths is shown in Figure 2. We used these evidence strength ratings in the design and analyses of Experiments 2 and 3.

3 Experiment E.2: production

Next, we evaluated speakers' intuitions in a forced production task, testing how likely they are to use a particular form to communicate their belief about q when given different pieces of evidence.[3]

[2]This experiment can be viewed here.
[3]This experiment can be viewed here.

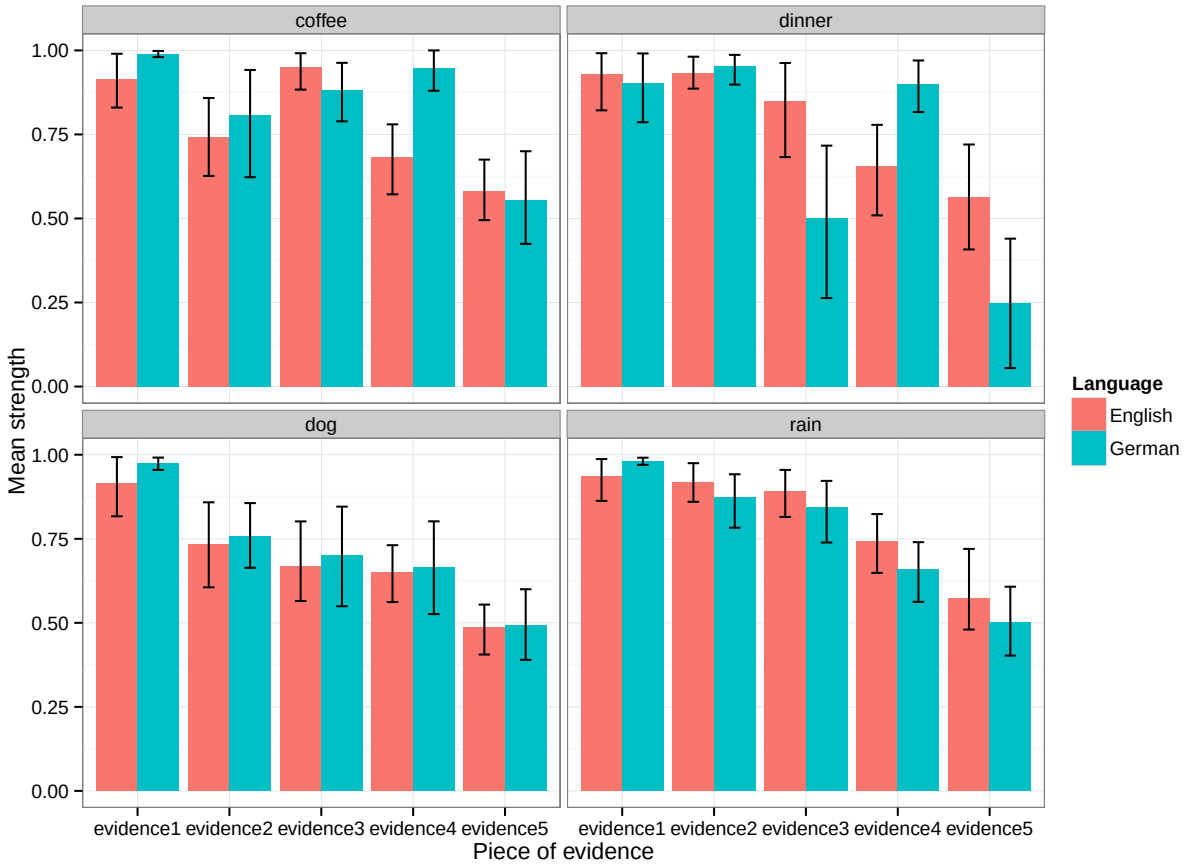

Figure 2: Mean by-item evidence strength for English and German, by domain. Error bars are bootstrapped 95% confidence intervals.THIS IS MOSTLY JUST FOR YOU GUYS TO HAVE A LOOK, unless you think this sort of plot would be useful to include. Basically, there are

3.1 Methods

3.1.1 Participants

We recruited 40 participants from Amazon's Mechanical Turk. Participants were compensated with a small payment.

3.1.2 Materials and procedure

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 possible utterances to describe the situation to a friend. On each trial, they first saw a context sentence which varied by domain, e.g., "Imagine that you are sitting in a room." Next, they were presented with a piece of evidence, e.g., "Earlier today, you had seen dark clouds in the sky." Finally, each participant saw the same question: "Given what you know, what do you say to a friend who is sitting in a windowless room down the hall?" They then chose one of four possible utterances by checking a radio button, e.g., "It's raining", "It must be raining", "It's probably raining", "It might be raining". Across domains, each choice was between utterances of the forms shown in (5).

(5) *Abstract form of utterance choices:*
   a. *q* (bare)
   b. *must q* (must)
   c. *probably q* (probably)
   d. *might q* (might)

Each participant completed 12 trials, three per domain. For each participant and domain, three pieces of evidence were randomly sampled from the set of five. Trial order was randomized, as was the order of utterance options.

3.2 Results and discussion

The overall distribution of utterance choices is shown in Figure 3. The bare and *might* forms are used most frequently, with both *must* and *probably* being chosen at only half the rate. The main question of interest in production is whether the choice of form to communicate about q depends on the strength of the evidence for q. Indeed it does: Figure 4 shows the mean strength of the evidence (as elicited in Exp. E.1) that participants were given as a function of the utterance they ultimately chose. In order to evaluate the effect of evidence strength on utterance choice, we conducted a mixed-effects linear regression predicting evidence strength from a dummy-coded predictor for utterance choice (with *must* as reference level) as well as random by-participant and by-item intercepts. Evidence strength was greater when the bare form was produced than when *must q* was produced (β = .11, SE = .01, t = 7.58, $p < .0001$) and smaller when *might q* was produced (β = -.13, SE = .01, t = -8.99, $p < .0001$). There was no difference in evidence strength between *must q* and *probably q* (β = -.01, SE = .02, t = -.73, $p < .47$). [jd: this may seem weird because it's exactly the other way round that we're interested in – ie how likely is someone to produce one of the four utterances as a function of evidence strength? but i couldn't figure out a good way of doing that, which i'm happy to elaborate on in person. what do you think about this way of doing it?].

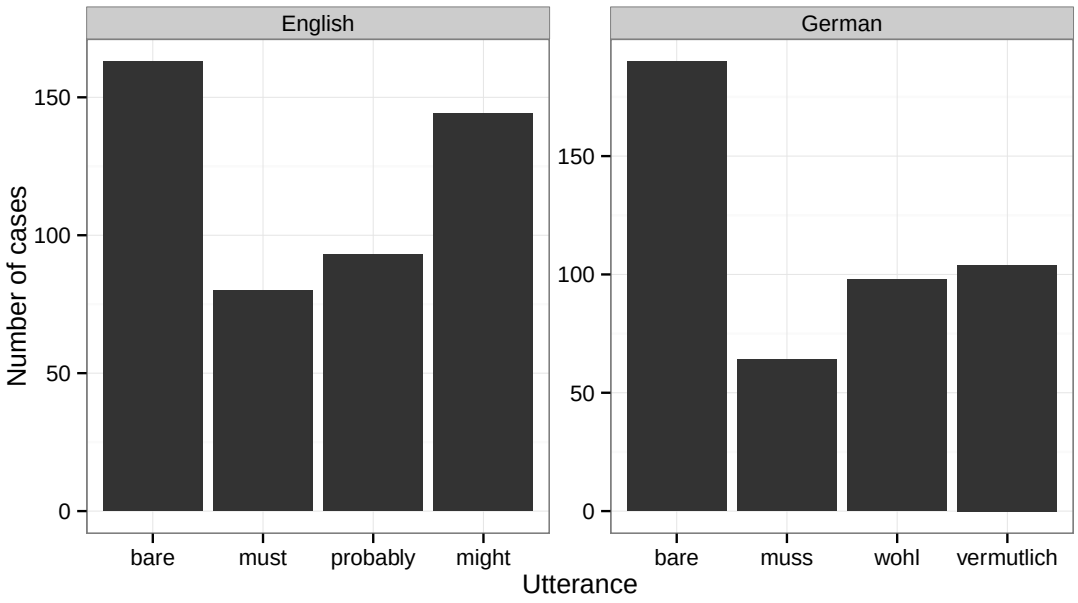

Figure 3: Histogram of utterance choice for English (Exp. E.2, left) and German (Exp. G.2, right).

Figure 4: Mean strength of evidence given when using each utterance, for English (Exp. E.2, left) and German (Exp. G.2, right). Error bars indicate bootstrapped 95% confidence intervals.

4 Experiment E.3a: comprehension (listener belief)

We next tested the other side of the communicative coin: depending on the utterance u used to communicate about q, how strong is listeners' resulting belief in q, and what do they believe to be the strength of the evidence the speaker was in possession of when producing u?[4]

4.1 Methods

4.1.1 Participants

We recruited 60 participants through Amazon's Mechanical Turk. Participants were compensated with a small payment.

4.1.2 Materials and procedure

Participants were presented with an utterance (e.g., "It must be raining") and asked a) to rate the probability of the state of affairs q (e.g., it is raining); and b) to select one out of five pieces of evidence that the speaker had about q in making their utterance. On each trial, participants first saw two context sentences: "You are in a windowless room. Your friend X walks in and says:", where "X" was a randomly generated name.[5] Participants then saw one of the utterances from Exp. E.2 that "X" produced, e.g., "It must be raining". They were asked "How likely do you think it is that it is raining?" and adjusted a slider with endpoints labeled "impossible" and "certain" in response. Once they thus submitted their belief in q, the five potential pieces of evidence previously used in Exps. E.1 and E.2 were shown and participants were asked to choose one by clicking a radio button in response to the question "How do you think X knows about the rain?"

Participants provided one set of judgments for each domain, resulting in four trials per participant. Each participant saw each type of utterance (bare, must, probably, might) across trials. Utterance types were randomly distributed across domains. Trial order was randomized, as was the order in which pieces of evidence were displayed.

4.2 Results and discussion

Two questions are of interest: first, does the probability of listener belief in q vary as a function of the observed utterance? Second, does the strength of the evidence for q inferred to be available to the speaker vary as a function of the observed utterance? To address the first question, we conducted a mixed effects linear regression predicting degree of belief in q from a dummy-coded utterance predictor with *must* as reference level. The model included random by-participant and by-item intercepts. Fig. 5 shows mean probability of listener belief in q by utterance: participants believed q was more likely after observing the bare utterance than after observing the *must* utterance (β=.24, SE=0.03, t=9.1, p<0.0001). In contrast, they believed q was less likely after observing *might* q (β=-.09, SE=0.03, t=-3.44, p<0.0008). There was no difference in resulting listener belief between *must* q and *probably* q (β=-.04, SE=0.03, t=-1.6, p<0.12). These results mirror the evidence strength effects found in production (Exp. E.2).

[4] This experiment can be viewed here.

[5] This was done to discourage effects of inferences about speaker-specific language use on interpretation.

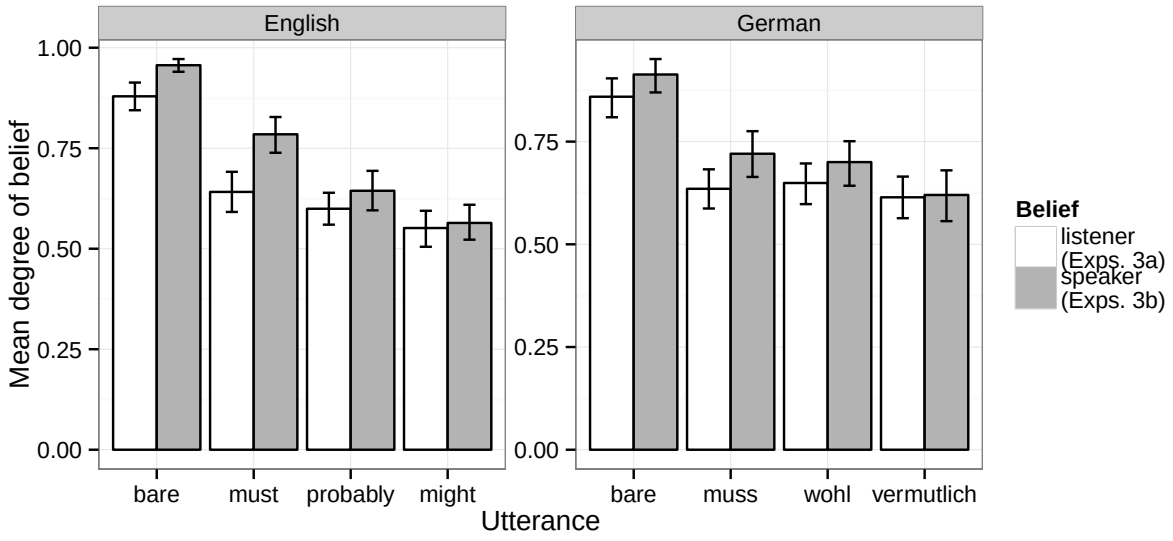

Figure 5: Mean probability of listener and speaker belief in q by utterance for English (left, Exps. E.3a and E.3b) and German (right, Exps. G.3a and G.3b). Error bars indicate 95% bootstrapped confidence intervals.

To address whether inferred speaker evidence strength mirrors production, we conducted another mixed effects linear regression, predicting inferred strength of evidence for q from a dummy-coded utterance predictor with *must* as reference level. The model included random by-participant and by-item intercepts. Figure 6 shows mean evidence strength ascribed to speakers by utterance: interestingly, participants inferred stronger evidence was available to the speaker after observing the bare utterance than *must* q (β=.08, SE=.02, t=3.74, $p<$ .0003), but inferred evidence strength was no different for *probably* q (β=.01, SE=0.02, t=.55, $p<$ .59) or *might* q (β=-.02, SE=0.02, t=-.89, $p<$ .38).

5 Experiment E.3b: comprehension (speaker commitment)

Exp. 3a tested listener beliefs in q as a function of the observed utterance. A related, but potentially orthogonal dimension is the commitment that listeners ascribe to speakers as a basis for producing a particular utterance. For example, a particular utterance may lead the listener to infer that the speaker is highly committed to q, while nevertheless not instilling the same degree of belief in q in the listener. In fact, epistemic *must* has been claimed to function like this: vFG claim that maximal speaker commitment is necessary for the use of epistemic *must*, just as in the use of the bare form; yet in comprehension the interpretation of *must* q is weaker than that of bare q. Exp. E.3b thus tested the degree of belief in q that listeners ascribe to *speakers* depending on the utterance the speaker produced.[6]

[6]This experiment can be viewed here.

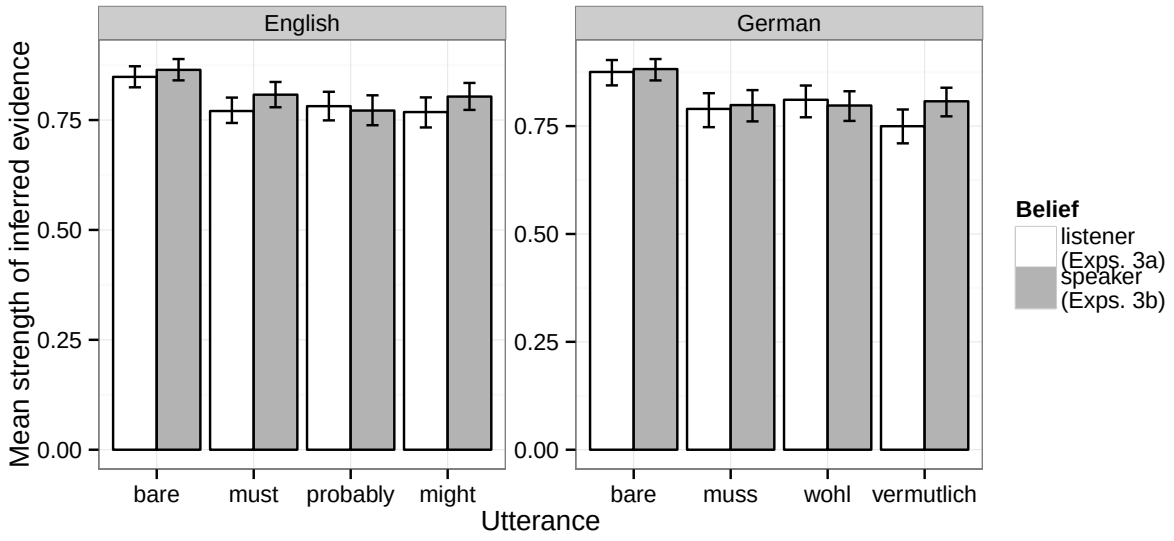

Figure 6: Mean inferred evidence strength by utterance for English (left, Exps. E.3a and E.3b) and German (right, Exps. G.3a and G.3b). Error bars indicate 95% bootstrapped confidence intervals.

5.1 Methods

5.1.1 Participants

We recruited 60 participants through Amazon's Mechanical Turk. Participants were compensated with a small payment.

5.1.2 Materials and procedure

The design, procedure, and materials were identical to those of Exp. 3a with the exception of the dependent measure: instead of asking participants how likely they thought that q, they instead answered the question "Does X think that it's raining?" by adjusting a slider on a scale with endpoints labeled "Definitely not" and "Definitely".

5.2 Results and discussion

As in Exp. E.3a, two questions are of interest: first, does the probability of belief in q – this time, as ascribed to the speaker rather than as the result of the listener's interpretation – vary as a function of the observed utterance? Second, does the strength of the evidence for q inferred to be available to the speaker vary as a function of the observed utterance? To address the first question, we conducted a mixed effects linear regression predicting degree of belief in q from a dummy-coded utterance predictor with *must* as reference level. The model included random by-participant and by-item intercepts. Fig. 5 shows mean probability of ascribed speaker belief in q by utterance: participants believed the speaker was more likely to believe q after observing the bare utterance than after observing the *must* utterance (β=.18, SE=0.03, t=6.59, p<0.0001). In contrast, they believed the speaker was less likely to believe q if they produced *probably* q (β=-.14, SE=0.03, t=-5.28, p<0.0001) or *might* q (β=-.22, SE=0.03, t=-8.36, p<0.0001).

These results mirror the effects found in Exp. E.3a, with the exception that all utterances led to differences in ascribed speaker commitment. Interestingly, the strength of the belief that participants attributed to speakers was stronger than their own resulting belief. This was borne out statistically in a model that was applied to both the listener and speaker belief datasets. This model was identical to that just reported, but additionally allowed for a dummy-coded belief holder predictor (listener vs. speaker) to interact with utterance. There was a clear main effect of belief holder, such that the belief ascribed to speakers was stronger than that held by listener participants (β=.14, SE=0.03, t=4.7, p<0.0001).

As in Exp. E.3a, to address whether inferred speaker evidence strength mirrors production, we conducted another mixed effects linear regression, predicting inferred strength of evidence for q from a dummy-coded utterance predictor with *must* as reference level. The model included random by-participant and by-item intercepts. Figure 6 shows mean evidence strength ascribed to speakers by utterance: again, participants inferred stronger evidence was available to the speaker after observing the bare utterance than *must q* (β=.06, SE=.02, t=3.2, p< .002), but inferred evidence strength was no different for *probably q* (β=-.02, SE=0.02, t=-1.11, p< .27) or *might q* (β=-.01, SE=0.02, t=-.49, p< .63).

Allowing this model to interact with a belief holder predictor and applying it to simultaneously to the Exp. E.3a dataset yields no main effect of belief holder (β=.03, SE=.02, t=1.53, p< .13) – this is unsurprising, given that this aspect of the dependent measure was identical across experiments.

6 Experiment G.1: evidence strength

In this experiment we collected estimates of evidence strength for the same pieces of evidence as in E.1. The experiment was identical to E.1 with the exception that it was conducted in German. [7]

6.1 Methods

6.1.1 Participants

40 participants were recruited through Clickworker's crowd-sourcing service, and were compensated for their participation.

6.1.2 Materials and procedure

The procedure was identical to that of Exp. E.1. All materials were translated into German. See Appendix A for the full list of stimuli.

6.2 Results and discussion

We obtained between 3 and 14 ratings for each piece of evidence. Participants' estimates of evidence strength are shown alongside those of the English-speaking participants in Figure 2. To test whether the English and German distributions of strength ratings differed, we conducted a mixed-effects linear regression predicting evidence strength rating from a dummy-coded fixed effect of language (with English as reference level) as well as by-participant and by-item random intercepts and by-item random slopes for language. The effect of language did not reach significance (β = -.01, SE

[7] This experiment can be viewed here.

= .03, t = -.22, $p < .83$), suggesting that the two populations did not differ in their estimates of evidence strength.

7 Experiment G.2: production

Next, we evaluated speakers' intuitions in a forced production task, testing how likely they are to use a particular form to communicate their belief about q when given different pieces of evidence. The experiment was identical to E.2 with the exception that it was conducted in German and contained slightly different utterance choices.[8]

7.1 Methods

7.1.1 Participants

We recruited 40 participants on the German crowd-sourcing service Clickworker. Participants were compensated with a small payment.

7.1.2 Materials and procedure

The procedure was identical to that of Exp. E.2. As for participants' utterance choices, we included bare q form and *must* q as in Exp. E.2, but instead of the modals *probably* and *might*, we included the modal adverbial *vermutlich* (English *presumably*) and the discourse particle *wohl*. issue of translating items into past perfect

7.2 Results and discussion

The overall distribution of utterance choices is shown in Figure 3 alongside the English results. The bare form is used most frequently, with the other forms only being chosen half as often, and indeed, *muss* q is dispreferred in general. We are again interested in whether the choice of form to communicate about q depends on the strength of the evidence for q: mean strength of the evidence (as elicited in Exp. E.1) that participants were given as a function of the utterance they ultimately chose is visualized inFigure 4. In order to evaluate the effect of evidence strength on utterance choice, we conducted a mixed-effects linear regression predicting evidence strength from a dummy-coded predictor for utterance choice (with *must* as reference level) as well as random by-participant and by-item intercepts. Mirroring the English result, evidence strength was greater when the bare form was produced than when *muss* q was produced (β = .12, SE = .03, t = 4.78, $p < .0001$). In addition, evidence strength was smaller when *vermutlich* q was produced (β = -.09, SE = .03, t = -3.35, $p < .0009$). There was no difference in evidence strength between *muss* q and *wohl* q (β = -.02, SE = .03, t = -.67, $p < .51$).

8 Experiment G.3a: comprehension (listener belief)

We next tested the other side of the communicative coin: depending on the utterance u used to communicate about q, how strong is listeners' resulting belief in q, and what do they believe to be the strength of the evidence the speaker was in possession of when producing u? This experiment

[8] This experiment can be viewed here.

was identical to E.3a with the exception that it was conducted in German and included the slightly different set of utterance options used in Exp. G.2. [9]

8.1 Methods

8.1.1 Participants

We recruited 60 participants through the German crowd-sourcing service Clickworker. Participants were compensated with a small payment.

8.1.2 Materials and procedure

The procedure was identical to that of Exp. E.3a, but materials were presented in German and the utterances participants observed were the ones used in Exp. G.2.

8.2 Results and discussion

As in the English case, two questions are of interest: first, does the probability of listener belief in q vary as a function of the observed utterance? Second, does the strength of the evidence for q inferred to be available to the speaker vary as a function of the observed utterance? To address the first question, we conducted a mixed effects linear regression predicting degree of belief in q from a dummy-coded utterance predictor with *muss* as reference level. The model included random by-participant and by-item intercepts. Fig. 5 shows mean probability of listener belief in q by utterance alongside the English results: participants believed q was more likely after observing the bare utterance than after observing the *muss* utterance (β=.22, SE=0.03, t=7.5, p<0.0001). However, there were no differences in degree of belief in q between *muss* q and *vermutlich* q (β=-.02, SE=0.03, t=-.76, p<.46) nor between *muss* q and *wohl* q (β=.01, SE=0.03, t=.43, p<.67).

To address whether inferred speaker evidence strength mirrors production, we conducted another mixed effects linear regression, predicting inferred strength of evidence for q from a dummy-coded utterance predictor with *muss* as reference level. The model included random by-participant and by-item intercepts. Figure 6 shows mean evidence strength ascribed to speakers by utterance alongside the English results: as in the English case, participants inferred stronger evidence was available to the speaker after observing the bare utterance than *muss* q (β=.08, SE=.02, t=3.23, p< .002). In addition, they inferred that the available evidence must have been weaker upon observing *vermutlich* q (β=-.05, SE=.02, t=-2.1, p< .04), but inferred evidence strength was no different for *wohl* q (β=.001, SE=0.02, t=.07, p< .95).

9 Experiment G.3b: comprehension (speaker commitment)

This experiment was a German version of Exp. E.3b and tested the commitment that listeners ascribe to speakers after observing each of the four utterance types used in Exps. G.2 and G.3a.[10]

[9] This experiment can be viewed here.
[10] This experiment can be viewed here.

9.1 Methods

9.1.1 Participants

We recruited 60 participants through the German crowd-sourcing service Clickworker. Participants were compensated with a small payment.

9.1.2 Materials and procedure

The procedure was identical to that of Exp. E.3b but was conducted in German and used the slightly different set of utterances.

9.2 Results and discussion

As in Exp. G.3a, two questions are of interest: first, does the probability of belief in q – this time, as ascribed to the speaker rather than as the result of the listener's interpretation – vary as a function of the observed utterance? Second, does the strength of the evidence for q inferred to be available to the speaker vary as a function of the observed utterance? To address the first question, we conducted a mixed effects linear regression predicting degree of belief in q from a dummy-coded utterance predictor with *muss* as reference level. The model included random by-participant and by-item intercepts. Fig. 5 shows mean probability of ascribed speaker belief in q by utterance alongside the English results: participants believed the speaker was more likely to believe q after observing the bare utterance than after observing the *muss* utterance (β=.19, SE=0.03, t=5.79, $p<$ .0001). In contrast, they believed the speaker was less likely to believe q if they produced *vermutlich* q (β=-.1, SE=0.03, t=-3.01, p<0.004). There was no difference between *muss* q and *wohl* q (β=-.02, SE=0.03, t=-.61, $p<$ .55).

Conducting the analysis jointly on the listener (Exp. G.3a) and speaker (Exp. G.3b) belief data and allowing a belief holder predictor to interact with the utterance predictor again yields a main effect of belief holder such that the belief ascribed to speakers is stronger than the resulting belief in participants (β=.08, SE=0.04, t=2.23, $p<$ .03).

As for Exp. G.3a, we tested whether inferred speaker evidence strength mirrors production by conducting another mixed effects linear regression, predicting inferred strength of evidence for q from a dummy-coded utterance predictor with *muss* as reference level. The model included random by-participant and by-item intercepts. Figure 6 shows mean evidence strength ascribed to speakers by utterance alongside the English results; the qualitative results were identical to those in the English case. Participants inferred stronger evidence was available to the speaker after observing the bare utterance than *muss* q (β=.08, SE=.02, t=4.05, $p<$ .0001). However, evidence strength was not inferred to be any different after observing *vermutlich* q (β=-.0003, SE=.02, t=-.02, $p<$ .99) or *wohl* q (β=.008, SE=0.02, t=.36, $p<$ .72).

Allowing a belief holder predictor to interact with the utterance predictor when applied to both the listener and the speaker dataset yielded no effect of belief holder (β=.0008, SE=0.03, t=.04, $p<$ .97).

10 General discussion

XXX

11 Conclusion

XXX

A Pieces of evidence

This section lists, for each proposition q, the five pieces of evidence that were used throughout all experiments.

A.1 It's raining. / Es hat geregnet.

1. You look out the window and see raindrops falling from the sky.
   Sie sehen aus dem Fenster und beobachten, wie Regentropfen vom Himmel fallen.

2. You hear the sound of water dripping on the roof.
   Sie knnen hren, wie Wasser auf das Dach prasselt.

3. You check the weather report on the Internet, which says it is raining.
   Sie haben im Internet den Wetterbericht gelesen, in dem stand, dass es regnen wrde.

4. You see a person come in from outside with wet hair and wet clothes.
   Sie sehen, wie jemand mit nassen Haaren und durchnssten Kleidern von drauen hereinkommt.

5. Earlier today, you had seen dark clouds in the sky.
   Sie haben heute Vormittag dunkle Wolken am Himmel gesehen.

A.2 The coffee is cold. / Der Kaffee ist kalt geworden.

1. You take a sip of the coffee and feel that it is cold.
   Sie trinken einen Schluck Kaffee und stellen fest, dass er kalt ist

2. You touch the coffee cup and feel that it is cold.
   Sie berhren die Kaffeetasse und stellen fest, dass sie kalt ist.

3. You see that there is no steam coming from the coffee.
   Sie sehen, dass aus dem Kaffee kein Dampf aufsteigt.

4. You know that the coffee has been on the table for an hour.
   Sie wissen, dass der Kaffee seit einer Stunde auf dem Tisch steht.

5. You see that the cup isn't insulated.
   Sie sehen, dass die Tasse nicht isoliert ist.

A.3 Dinner is ready. / Das Abendessen ist fertig geworden.

1. You just prepared dinner and set it out on the table.
   Sie haben gerade das Abendessen zubereitet und auf den Tisch gestellt

2. Your spouse tells you that dinner is ready.
   Ihr/e Partner/in sagt, dass das Abendessen fertig ist.

3. Dinner is usually ready at around 6pm. You look at the clock and it is 6pm.
   Sie wissen, dass das Abendessen normalerweise um 18 Uhr fertig ist. Ein Blick auf die Uhr zeigt, dass es gerade 18 Uhr ist.

4. You smell food coming from the dining room.
   Sie vernehmen den Geruch von Essen, der aus dem Esszimmer kommt.

5. You're hungry.
   Sie haben Hunger.

A.4 The neighbor's dog is barking. / Der Nachbarshund hat gebellt.

1. You look outside and see Fluffy, the neighbor's dog, standing on the porch and barking.
   Sie schauen aus dem Fenster und sehen Struppi, den Hund der Nachbarn, wie er am Zaun steht und bellt.

2. You hear the sound of a dog barking.
   Sie hren einen Hund bellen.

3. You are listening to music with your earphones. You know that your neighbor's dog often barks in the evening.
   Sie haben Kopfhrer auf und hren Musik, wissen aber, dass der Hund der Nachbarn abends oft bellt.

4. You are listening to music with your earphones. You look out the window and see that the mailman has just arrived at your neighbor's doorstep, when all of a sudden he jumps back.
   Sie haben Kopfhrer auf und hren Musik, sehen aber aus dem Fenster und beobachten, wie der Postbote vor der Nachbarstr einen Satz nach hinten macht.

5. Your neighbor just got a new dog.
   Sie wissen, dass sich die Nachbarn gerade einen Hund angeschafft haben.

B Acknowledgments

References

Bergen, L., Levy, R., & Goodman, N. D. (2014). *Pragmatic reasoning through semantic inference.* (Unpublished MS, Available online at http://web.mit.edu/bergen/www/papers/BergenLevyGoodman2014.pdf)

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998-998.

Grice, P. (1989). *Studies in the way or words.* Cambridge, MA: Harvard University Press.

Karttunen, L. (1972). *Possible* and *must.* In J. Kimball (Ed.), *Syntax and semantics* (Vol. 1, p. 1-20). New York: Academic Press.

Kratzer, A. (1991). Modality. In A. von Stechow & D. Wunderlich (Eds.), *Semantics: An international handbook of contemporary research* (p. 639-650). Berlin: de Gruyter.

Lassiter, D. (2011). *Measurement and modality: The scalar basis of modal semantics.* Unpublished doctoral dissertation, New York University.

Lassiter, D. (*to appear*). The weakness of *must*: In defense of a mantra. In T. Snider (Ed.), *Proceedings of SALT 24.*

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In T. Snider (Ed.), *Proceedings of SALT 23* (p. 587-610). CLC Publications.

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* Cambridge, MA: MIT Press.

Veltman, F. (1985). *Logics for conditionals.* Unpublished doctoral dissertation, University of Amsterdam.

von Fintel, K., & Gillies, A. S. (2010). *Must* ... stay ... strong! *Natural Language Semantics*, *18*, 351-383.