Definitely, maybe:
A new experimental paradigm for investigating the pragmatics of evidential devices across languages

Abstract. We present a new experimental paradigm for investigating lexical expressions that convey different strengths of speaker commitment. Specifically, we compare different evidential contexts for using modal devices, epistemic discourse particles, and statements with no evidential markers at all, examining the extent to which listeners' interpretations of certain types of evidential words and their judgments about speaker commitment differ in strength. We also probe speakers' production preferences for these different devices under varying evidential circumstances. The results of our experiments shed new light on distinctions and controversies that play a key role in the current theoretical literature on the semantics and pragmatics of modals and discourse particles. Our paradigm thus contributes to a domain of experimental research on evidential expressions that is only just taking shape at the crossroads of theoretical semantics/pragmatics and psycholinguistics; we provide a starting point for approaching theoretical debates on the nature of evidential expressions from an experimental and context-oriented perspective.

Keywords: discourse particles; English; evidentials; German; modals; psycholinguistics

1 Introduction

Recent years have seen a rapid development in typological and empirical semantic research on evidential expressions across languages (see, e.g., Korotkova 2016; Matthewson 2012; Murray 2017). Within this growing body of research, we observe a recent trend towards experimental attempts to systematically explore different aspects of the use of evidential devices (see Knobe & Yalcin 2014; Lassiter 2016; Ünal & Papafragou 2018). The series of experiments we present in this paper contributes to this type of empirical research by providing a new experimental paradigm for investigating lexical expressions that convey different strengths of speaker commitment. In particular, we introduce a methodology for exploring the impact of different evidential circumstances on the use of modal evidentials, epistemic discourse particles, and statements with no evidential markers at all. In the process, we indicate how these case studies can profitably be linked to issues and controversies found in the current theoretical literature. Before turning to these issues in more detail, let us briefly sketch the evidential devices we are concerned with in this paper.

When speakers are certain about some fact expressed by a proposition *p* (e.g., that it is raining), they are likely to communicate this fact with a simple declarative utterance, as in (1a). When they want to convey epistemic uncertainty, they have different devices for doing so: adding a modal adverbial, such as *probably* (1b), or using modal verbs as in (1c) and (1d). Either option leaves open the possibility that *p* is not true.

(1) a. It is raining.
b. It is probably raining.
c. It might be raining.
d. It must be raining.

Other languages have more varied ways of conveying different degrees of speaker commitment (for an overview, see Aikhenvald 2004). Even between closely related languages, we observe striking differences. German, for instance, is an interesting comparison case to English. While many devices for expressing speaker certainty do not differ from English---examples (2a)-(2c) are the German equivalents of (1a), (1b), and (1d)---German possesses a lexical inventory of so-called 'discourse particles', which English lacks. Consider (2d), which is an example involving the epistemic discourse particle *wohl* (lit. 'well'); see Zimmermann (2004) and many others:

(2) a. Es regnet.
it is.raining
'It is raining.'
b. Es regnet vermutlich.
it is.raining probably
'It is probably raining.'
c. Es muss regnen.
it must be.raining
intended reading: 'It must be raining.'
d. Es regnet wohl.
it is.raining PRT
'It is PRT raining (I guess).'

Discourse particles like *wohl* organize the discourse by conveying the epistemic states of both the speaker and the listener. In our case, the particle *wohl* communicates that the speaker merely assumes that the propositional content is true, and thus *wohl* lexically encodes weakened speaker commitment.

In our series of experiments below, we restrict ourselves to the epistemic/evidential words in English and German that are given in (1) and (2). In doing so, we hope to provide a focused set of experimental studies on the expression of speaker commitment that can be seen as an articulated road map for future experimental work in this domain. We are aware that our choice of evidential devices represents only a very small portion of cross-linguistic options to express different degrees of speaker commitment. However, in the next two sections, we will clarify why we chose these particular cases in English and German. We will point out to what extent experimental data on them might be interesting for claims and controversies that can be found in the current theoretical literature, and we will argue that our experiments on different degrees of speaker commitment in the domain of these evidential devices address obvious blind spots in the theoretical debates.

1.1 Issue I: Evidential circumstances for using epistemic *must*

The weak reading of *must* in (1d) above has puzzled linguistic theory for a long time already. *Must* serves as a strong modal of necessity, so its interpretation regarding speaker commitment should not be weaker than the plain statement without a modal in (1a); see Karttunen (1972) and subsequent work. More formally, if we assume a quantificational treatment of modality, necessity modals like *must* correspond to universal quantifiers over possible worlds (e.g., Kratzer 1991). The necessity modals assert that in every (relevant) possible world, the proposition *p* holds. Accordingly, epistemic *must* in (1d) should assert that in every world compatible with the speaker's knowledge, it is raining. Given that

knowledge corresponds to justified true belief (i.e., that which is known cannot be otherwise), from (1d) it necessarily follows that it is raining (i.e., it is not possible that it is not raining). At issue, then, is the failed inference from (1d) to (1a): how could *must p* not entail *p*? If it is *necessarily* raining, then surely it is raining. However, it appears that talking about what is necessarily the case commits speakers to less than does talking about what is *actually* the case.

To account for this puzzle, several different approaches have been proposed. One prominent account is the two-part theory of epistemic *must* from von Fintel & Gillies (2010, 2016). The authors propose that the necessity modal *must* indeed stays 'strong', with *must p* entailing *p*. At the same time, *must* introduces a lexical presupposition according to which *p* is known as a result of indirect inference. Let us briefly illustrate this idea.

Von Fintel & Gillies (2010) object to the claim that *must* statements like (3a) are 'weaker' than those made by the prejacent alone, (3b):

(3) a. It must be raining.
b. It is raining.

They argue that theories postulating that *must* serves to make weak claims in cases such as (3a) confuse the notion of indirectness with 'a feeling of weakness'. In support of this distinction, von Fintel & Gillies provide examples such as the following (von Fintel & Gillies 2010: 362):

(4) CONTEXT: Chris has lost her ball, but she *knows* with full certainty that it is either in Box A or B or C. She says:
The ball is in A or in B or in C.
It is not in A. ... It is not in B.
So, it must be in C.

The point of this example is that Chris knows the ball must be in one of the boxes. After excluding two of them, Chris knows conclusively that the ball is in box C. Since her knowledge is a matter of logical deduction, it can be considered indirect knowledge. Crucially, the final statement *So, it must be in C* is then not weak at all because Chris knows with full certainty that the ball is in C. Using the modal *must* here thus simply expresses the indirect nature of Chris' knowledge.

Given such examples, one should reconsider the evidential contexts that were often cited to claim that *must* is weak. Consider the following example from von Fintel & Gillies (2010: 353):

(5) [Seeing the pouring rain]
a. It's raining.
b. ??It must be raining.

Von Fintel & Gillies (2010) argue that sentences like (5b) are not infelicitous in a context like (5) because they are too weak. Rather, according to their approach, the infelicity of (5b) is due to the fact that the statement is marked as being based on indirect evidence in a context that clearly features direct evidence. Accordingly, von Fintel & Gillies (2010) place a premium on the nature (i.e., direct vs. indirect) of the evidence.

Already we see that the evidential circumstances are crucial for understanding the meaning of epistemic *must* (see also Matthewson 2015). All existing approaches on epistemic *must* agree on this point. The question thus shifts to where one ought to locate this

evidential component in the meaning of *must*. There are several accounts that disagree with the presuppositional status of this evidential component proposed by von Fintel & Gillies (2010, 2016). In particular, some argue that it should rather be considered a conventional implicature (e.g., Salmon 2011), while others claim that it can be accounted for by Gricean reasoning. Under this latter style of approach, *must* is yet again treated as weak, with the evidential component of its meaning derived as a conversational implicature (Lassiter 2014, 2016; Giannakidou & Mari 2016; Goodhue 2016; Mandelkern 2017).

The experimental data we present below cannot decide between all of these accounts. However, if epistemic *must* means something like 'it follows from the evidence that', we can raise the issue of what kind of evidence (let's say on a scale from first-hand observations to rather vague inference) exactly allows or even favors the use of epistemic *must*. As far as we know, no systematic attempt has so far been made to test for different degrees of evidence strength in this context. Before reporting on the relevant experiments, let us now turn to a second empirical blind spot in the theoretical literature on evidential devices.

1.2 Issue II: Evidential circumstances for using epistemic discourse particles

As already pointed out in example (2) above, some languages additionally feature discourse particles to express epistemic meaning components. In this domain, most work has been carried out in formal semantics (see Grosz 2016 for a recent overview), and to date there is very little experimental work (see, however, Döring & Repp 2016; Dörre & Trotzke 2017).

In what follows, we will be concerned with the German discourse particle *wohl* (lit. 'well'). According to the literature, *wohl* in (6a) amounts to a quasi-synonym of epistemic *muss* ('must') (6b) and modal adverbs such as *vermutlich* ('probably') in (6c) (Grosz 2017; Zimmermann 2004); the particle *wohl* can thus be compared to our English data in interesting ways.

(6) a. Es regnet wohl.
it is.raining PRT
'It is PRT raining (I guess).'
b. Es muss regnen.
it must be.raining
intended reading: 'It must be raining.'
c. Es regnet vermutlich.
it is.raining probably
'It is probably raining.'

Note that particles like *wohl* are often viewed as a special type of adverbs from a structural perspective (e.g., Cardinaletti 2011). The claim in many formal approaches is that several properties of discourse particles derive from more general constraints on sentence adverbs (both in the syntax and in the semantics). For instance, both particles and sentence adverbs such as *probably* cannot occur in the surface scope of sentential negation; examples adapted from Grosz (2016: 4):

(7) a. Das ist {**wohl**} nicht {* **wohl**} seine Schuld.
that is PRT not PRT his fault
'That PRT isn't his fault (I guess).'
b. She {**probably**} hasn't {***probably**} left.

However, it is well known that there are several clear-cut syntactic distinctions between sentence adverbs and discourse particles. One of the most striking structural differences is that adverbs (8a) but not discourse particles (8b) can appear in front of the finite verb in matrix clauses:

(8) a. Vermutlich ist das nicht seine Schuld.
probably is that not his fault
'Probably, that isn't his fault.'
b. * Wohl ist das nicht seine Schuld.
PRT is that not his fault
'That isn't his fault (I guess).'

On the semantic side, however, it is not so clear to what extent *wohl* differs from its modal counterparts. There are different views on how *wohl* might differ from adverbs and epistemic *must* concerning scope-taking in question formation and structured propositions (e.g., Zimmermann 2008). Most of these differences are based on the assumption that *wohl* is not part of the truth-conditional content of the clause, whereas the evidential component of *must* and also the epistemic contribution by adverbs contribute to truth-conditional content. However, this distinction cannot be taken for granted, given what we said above about epistemic *must* and given the evidence for non-truth-functional views on higher adverbs discussed by Ernst (2007) and many others. Taken together, the affinities between discourse particles and other modal devices for expressing similar meanings indicates that the semantic boundary between these different evidential devices is perhaps fluid and that proposed distinctions rely on subtle theoretical and empirical issues.

In what follows, we will therefore investigate the extent to which *wohl* differs from its modal counterparts in its compatibility with different evidential contexts. Although all evidential devices (epistemic *must*, adverbs, and discourse particles) modify the speaker's commitment to a proposition, no study to date has explored whether these (at first sight synonymous) expressions differ in their compatibility with different evidential circumstances.

In the next few sections, we will illustrate our new experimental paradigm for investigating evidential words by comparing the theoretically-challenging devices introduced in Section 1.1 and 1.2 – epistemic *must* and epistemic discourse particles – with other devices such as English *might* and *probably*, German *vermutlich*, and the unmarked bare form. In doing so, we will address the following questions: (i) Under which evidential circumstances do speakers prefer to use which evidential devices? Put differently, how do speakers use the various evidential devices? and (ii) Do listeners ascribe different strengths of speaker commitment to the use of these various linguistic devices? In other words, how do listeners interpret speakers' use of these devices? Before directly addressing these questions, we first introduce our experimental materials and obtain estimates of evidence strength. These data will then serve as the basis for the analyses that follow.

2 Experiment 1: Evidence strength

To investigate the role of evidence strength in the production and comprehension of evidential devices, we must first generate a set of pieces of evidence that vary in how they

provide support for a proposition *p*. Thus, in this section we report on an experiment that collected estimates of evidence strength with the goal of norming evidence types for a variety of characteristics that will serve as the basis for the studies to be presented below.[1]

2.1 Methods

2.1.1 Participants

For the English version of the experiment, 40 native English speakers were recruited through Amazon's Mechanical Turk crowd-sourcing service. For the German version, 40 German native speakers were recruited through Clickworker's crowd-sourcing service. Both groups were compensated for their participation.

2.1.2 Materials and procedure

Participants rated the probability of a state of affairs *p* given a piece of evidence *e* by adjusting a slider on a scale with endpoints labeled "impossible" and "absolutely certain."[2] On each trial, participants first saw the following context sentence: "Imagine that you are at home." Then the evidence *e* for *p* was shown, for example, "Dinner is usually ready around 6pm. You look at the clock and it is 6pm." Finally, participants were asked about the probability of *p*, "How likely is it that dinner is ready?" and adjusted the slider accordingly. There were four different states of affairs *p* that appeared in the "How likely is it that *p*?" frame:

(9) a. it is raining
    b. the coffee is cold
    c. dinner is ready
    d. the neighbor's dog is barking

For each *p*, each participant evaluated one of five possible pieces of evidence, resulting in four trials per participant. Trial order was randomized. For the German version, the procedure was identical; all materials were translated into German. See Appendix A for the full list of stimuli.

2.2 Results and discussion

We obtained between 3 and 14 strength 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 *p* given *e*, which we will also refer to as *evidence strength*. Strength of each piece of evidence in both the English and German task is shown in Figure

[1] The English version of this experiment can be viewed here. The German version can be viewed here.

[2] Pieces of evidence were generated through a separate English free production paradigm. A speaker's description of some state of affairs *p* was given to participants and they were asked to provide a free response explanation of how the speaker knew about *p*. This experiment can be viewed here. Similar responses (e.g., "he can hear it" and "he can hear it on the roof") were grouped together. We selected explanations from among those most frequently generated while including at least one example of each type of evidence (direct, indirect, reportative). This procedure resulted in the final selection of five pieces of evidence per state of affairs. English materials were translated into German by authors 1, 2, and 4, who are native speakers of German.

1. We used these pieces of evidence in the design of Experiments 2 and 3; analyses employed evidence strength means (indicated by black dots in the Figure).

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

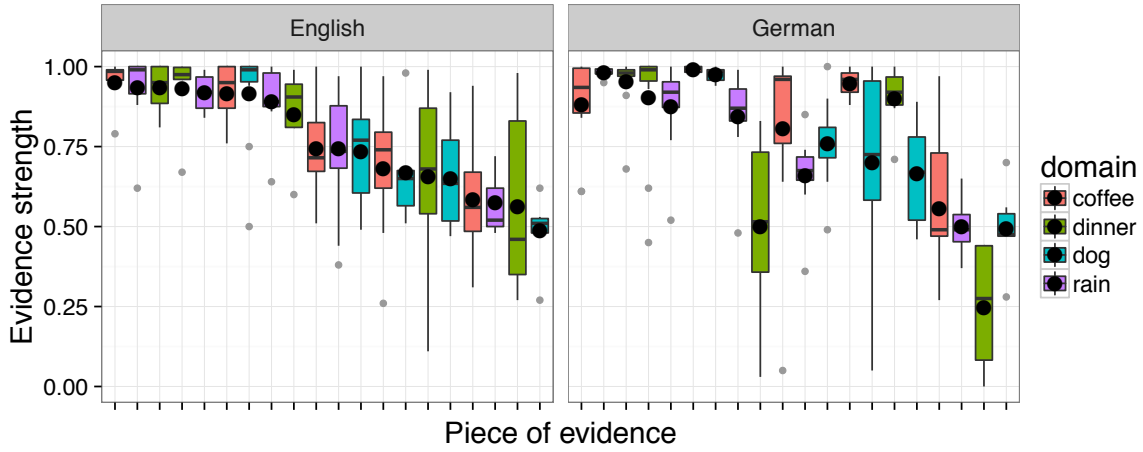

Figure 1: Boxplots of by-item evidence strength for English (left) and German (right). Pieces of evidence occur in the same order in the English and German panels and are sorted by the English means. Horizontal lines indicate medians, black dots indicate means. Boxes indicate the range into which 50% of the data fall. Whiskers extend to 1.5 times the interquartile range.

3 Experiment 2: Production

The following experiment addressed the first main question of interest: under what evidential circumstances do speakers use which evidential devices? In other words, what are the evidential use conditions for the cross-linguistic devices we focus on? We thus investigate the interaction between speakers' choice of closely-related evidential expressions and concrete scenarios. To this end, we evaluated speakers' intuitions in a forced production task, testing how likely they are to use a particular evidential device to communicate their belief about *p* when confronted with pieces of evidence that differ in how strongly they support *p*.[3] The German version was identical with the exception that it was conducted in German and contained slightly different utterance choices (explained below).[4]

3.1 Methods

3.1.1 Participants

[3] The English version of this experiment can be viewed here.
[4] The German version of this experiment can be viewed here.

For the English version, we recruited 40 participants from Amazon's Mechanical Turk. For the German version, we recruited 40 participants on the German crowd-sourcing service Clickworker. Participants were compensated for their participation.

3.1.2 Materials and procedure

Participants 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 saw 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"). Depending on the language of testing, possible utterances took the forms shown in (10) or (11); for German we included the bare *p* form and *must p* as in the English version, but instead of the modals *probably* and *might*, we included the modal adverbial *vermutlich* ('probably') and the discourse particle *wohl*. This design allows us to compare not only certain types of modals to the epistemic use of *must* (and to statements with no evidential markers at all), but also to compare both modals and *must* to discourse particles.

(10) *Form of English utterance choices:*
a. *p* (bare)
b. *must p* (must)
c. *probably p* (probably)
d. *might p* (might)

(11) *Form of German utterance choices:*
a. *p* (bare)
b. *muss p* (muss)
c. *wohl p* (wohl)
d. *vermutlich p* (vermutlich)

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 2. In both English and German, the bare form is used most frequently. In English, *might* is also used frequently, with both *must* and *probably* being chosen at only half the rate. A similar picture obtains in German, where *muss p* is generally dispreferred.

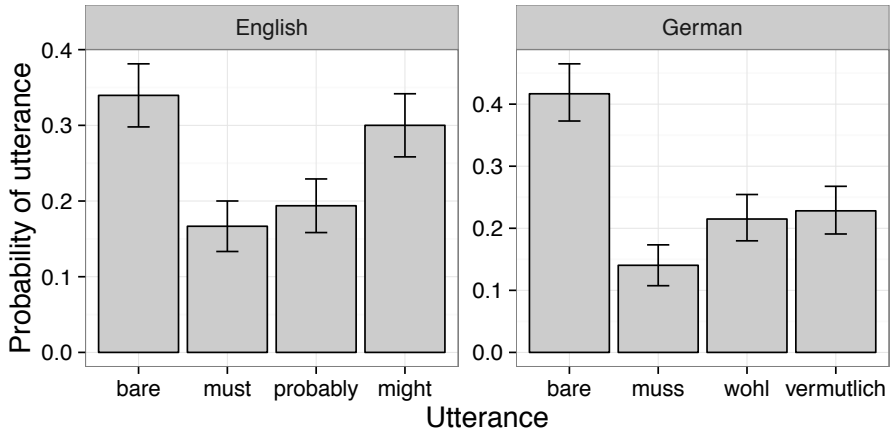

Figure 2: Probability of utterance choice for English (left) and German (right). Error bars indicate bootstrapped 95% confidence intervals.

The main question of interest is whether the choice of form to communicate about *p* depends on the strength of the evidence for *p*. Indeed, it does: Figure 3 shows the mean strength of the evidence (as elicited in Experiment 1) that participants were given as a function of the utterance they ultimately chose. Figure 4 shows the proportion with which each utterance was chosen as a function of evidence strength. In order to evaluate the effect of evidence strength on utterance choice, we conducted two separate mixed-effects linear regressions – one for the English data, one for the German data – predicting evidence strength from a dummy-coded predictor for utterance choice. In both cases, *must/muss* served as reference level. The models included random by-participant and by-item intercepts. Evidence strength was greater when the bare form was produced than when *must/muss q* was produced in both English (β = .11, *SE* = .01, *t* = 7.58, $p < .0001$) and German (β = .12, *SE* = .03, *t* = 4.78, $p < .0001$). In English, evidence was weaker when *might p* was produced (β = -.13, *SE* = .01, *t* = -8.99, $p < .0001$). We found no difference in evidence strength between *must p* and *probably p* (β = -.01, *SE* = .02, *t* = -.73, $p < .47$). In German, evidence was weaker when *vermutlich p* was produced (β = -.09, *SE* = .03, *t* = -3.35, $p < .0009$). We found no difference in evidence strength between *muss p* and *wohl p* (β = -.02, *SE* = .03, *t* = -.67, $p < .51$).

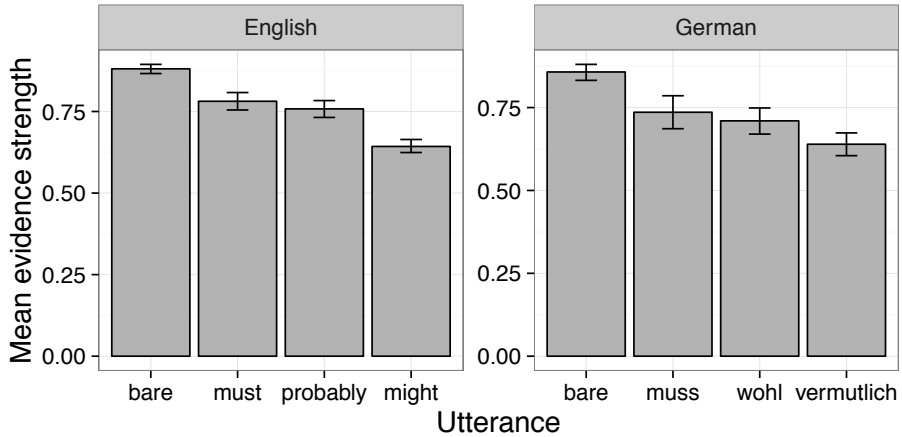

Figure 3: Mean strength of evidence given when using each utterance, for English (left) and German (right). Error bars indicate bootstrapped 95% confidence intervals.

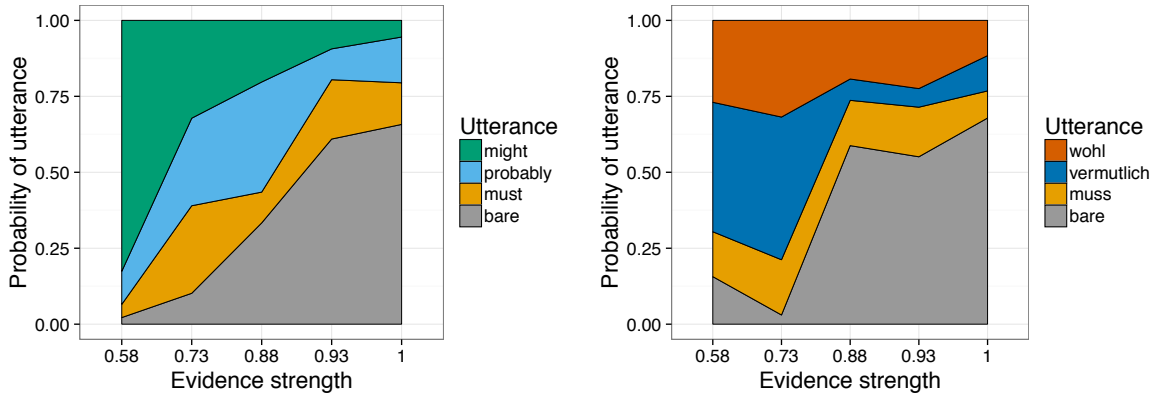

Figure 4: Probability of utterance choice as a function of evidence strength for English (left) and German (right). X-axis labels indicate the maximum evidence strength for the bin that the utterance proportion was computed over.

3.3 Discussion

We interpret these results as follows. First, bare utterances result from maximally strong evidence: when certainty of *p* is high, the claim that *p* may be made directly. As evidence strength decreases, speakers employ evidential devices that track evidence strength. In German, due to the lexical inventory of discourse particles, speakers have a choice of using epistemic adverbs such as *vermutlich*, epistemic *muss*, and particles like *wohl*. As Figure 3 shows, speakers tend to use *muss* and *wohl* instead of the respective adverb when the degree of evidence strength is higher. In other words, when investigating the dependence on evidence strength for *p*, we find that *muss* and *wohl* pattern together, in contrast to other modal means such as *vermutlich*. This is an interesting result given what we have discussed in Section 1.2 above, since our results suggest a meaning difference between discourse particles and otherwise synonymous adverbs in the domain of speaker commitment that has not been observed in the theoretical literature at all.

4 Experiment 3a: Comprehension (listener belief)

We next tested the flip side of the communicative coin: What are the inferences that listeners draw when observing the various evidential devices explored in Experiment 2? In particular, depending on the utterance *u* used to communicate about *p*, (i) how strong are listeners' resulting beliefs in *p*; (ii) what do they take the speaker to be committed to in uttering *u*; and (iii) what do they believe to be the strength of the evidence for *p* the speaker was in possession of when producing *u*? Experiment 3a addresses questions (i) and (iii) and Experiment 3b addresses questions (ii) and (iii). We first report Experiment 3a.[5] The English and German experiments were identical except for the language of testing and the target utterances presented to participants: English participants saw the set in (10), German participants the one in (11).

4.1 Methods

[5] The English version of this experiment can be viewed here and the German version here.

4.1.1 Participants

For the English version, we recruited 60 participants through Amazon's Mechanical Turk. For the German version, we recruited 60 participants through the German crowd-sourcing service Clickworker. Participants were compensated for their participation.

4.1.2 Materials and procedure

Participants were presented with an utterance *u* (e.g., "It must be raining") and asked to both rate the probability of the state of affairs *p* obtaining (e.g., it is raining) and to select one out of five pieces of evidence that the speaker most likely had about *p* in choosing the 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.[6] Participants then saw one of the utterances from Experiment 2 that "X" produced (e.g., "It must be raining"). They were then asked about the strength of their belief in *p*: "How likely do you think it is that it is raining?" and adjusted a slider with endpoints labeled "impossible" (coded as 0) and "certain" (coded as 1). Once they indicated their belief in *p*, the five potential pieces of evidence previously used in Experiments 1 and 2 were shown and participants were asked to choose the one the speaker likely had: "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 (English: bare, *must*, *probably*, *might*; German: bare, *muss*, *wohl*, *vermutlich*) 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 *p* vary as a function of the observed utterance? In other words, how are listeners' beliefs influenced by various evidential devices? Second, does the strength of the evidence for *p* 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 *p* from a dummy-coded utterance predictor with *must/muss* as reference level, separately for English and German. The model included random by-participant and by-item intercepts. Fig. 5 shows mean probability of listener belief in *p* by utterance: participants believed *p* was more likely after observing the bare utterance than after observing the *must* utterance in English (β=.24, *SE*=0.03, *t*=9.1, $p<0.0001$) and the *muss* utterance in German (β=.22, *SE*=0.03, *t*=7.5, $p<0.0001$). In contrast, in English they believed *p* was less likely after observing *might p* (β=-.09, *SE*=0.03, *t*=-3.44, $p<0.0008$). There was no difference in resulting listener belief between *must p* and *probably p* (β=-.04, *SE*=0.03, *t*=-1.6, $p<0.12$). And in German, there were no differences in degree of belief in *p* between *muss p* and *vermutlich p* (β=-.02, *SE*=0.03, *t*=-.76, $p<.46$), nor between *muss p* and *wohl p* (β=.01, *SE*=0.03, *t*=.43, $p<.67$).

[6] The naming of speakers was done to discourage effects of inferences about speaker-specific language use on interpretation.

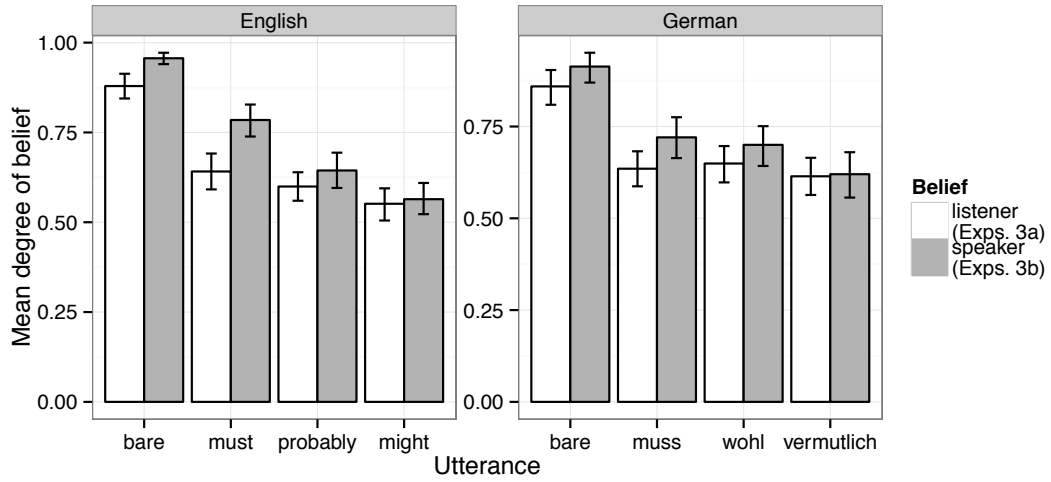

Figure 5: Mean probability of listener and speaker belief in p by utterance for English (left) and German (right). Error bars indicate 95% bootstrapped confidence intervals.

To address whether inferred speaker evidence strength mirrors the production results in Section 3, we conducted another mixed effects linear regression, this time predicting inferred strength of evidence for *p* from a dummy-coded utterance predictor with *must/muss* as reference level, separately for English and German. The model included random by-participant and by-item intercepts. Figure 6 shows mean evidence strength ascribed to speakers by utterance. Interestingly, in English, participants inferred stronger evidence was available to the speaker after observing the bare utterance than *must p* (β=.08, *SE*=.02, *t*=3.74, *p*< .0003), but inferred evidence strength was no different for *probably p* (β=.01, *SE*=0.02, *t*=.55, *p*< .59) or *might p* (β=-.02, *SE*=0.02, *t*=-.89, *p*< .38). Similarly in German, participants inferred stronger evidence was available to the speaker after observing the bare utterance than *muss p* (β=.08, *SE*=.02, *t*=3.23, *p*< .002). In addition, they inferred that the available evidence must have been weaker upon observing *vermutlich p* (β=-.05, *SE*=.02, *t*=-2.1, *p*< .04), but inferred evidence strength was no different for *wohl p* (β=.001, *SE*=0.02, *t*=.07, *p*< .95).

Taken together, the results of the current experiment on comprehension mirror those from production: bare utterances lead to the greatest degree of belief in *p* while indicating that the speaker had access to maximally strong evidence. Speaker belief and inferred evidence strength decrease with the epistemic modals; *must/muss* patterns with *probably*/*vermutlich*. Crucially, *muss* also patterns with the discourse particle *wohl*. We observed the same similarity in behavior between *muss* and *wohl* in the context of production in Experiment 2. In Section 6 below, we return to these effects (in both comprehension and production) in more detail.

5 Experiment 3b: Comprehension (speaker commitment)

Experiment 3a tested listener belief in *p* as a function of the observed utterance. A related dimension is the commitment to the truth of *p* that listeners ascribe to speakers. For example, a particular utterance may lead the listener to infer that the speaker is highly committed to (i.e., holds a strong belief in) *p*, while nevertheless not instilling the same degree of belief in *p* in the listener. In fact, epistemic *must* has been claimed to function like

this: von Fintel & Gillies (2010) 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 p* is weaker than that of bare *p*. That is, after hearing *must p*, listeners form beliefs in *p* that are weaker than what they take to be speakers' beliefs in *p*. Experiment 3b thus tested the degree of belief in *p* that listeners ascribe to *speakers* depending on the utterance the speaker produced.[7]

5.1 Methods

5.1.1 Participants

We recruited 60 English native speakers through Amazon's Mechanical Turk, and 60 German native speakers through Clickworker. Participants were compensated for their participation.

5.1.2 Materials and procedure

The design, procedure, and materials were identical to those of Experiment 3a with the exception of the dependent measure: instead of asking participants about their own strength of beliefs in *p*, they instead evaluated the speaker's belief in *p*: "Does X think that it's raining?" Participants indicated their response by adjusting a slider on a scale with endpoints labeled "Definitely not" (coded as 0) and "Definitely" (coded as 1).

5.2 Results and discussion

As in Experiment 3a, two questions are of interest: first, does the probability of belief in *p* – this time, as ascribed to the speaker rather than the listener's own belief – vary as a function of the observed utterance? Second, does the strength of the evidence for *p* 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 *p* from a dummy-coded utterance predictor with *must/muss* as reference level, separately for English and German. The model included random by-participant and by-item intercepts. Fig. 5 shows mean probability of ascribed speaker belief in *p* by utterance: participants believed the speaker was more likely to believe *p* after observing the bare utterance than after observing the *must* utterance in English (β=.18, *SE*=0.03, *t*=6.59, $p<0.0001$) or the corresponding *muss* utterance in German (β=.19, *SE*=0.03, *t*=5.79, $p< .0001$). In contrast, they believed the speaker was less likely to believe *p* if they produced *probably p* (β=- .14, *SE*=0.03, *t*=-5.28, $p<0.0001$) or *might p* (β=-.22, *SE*=0.03, *t*=-8.36, $p<0.0001$). In German, participants believed the speaker was less likely to believe *p* if they produced *vermutlich p* (β=-.1, *SE*=0.03, *t*=-3.01, $p<0.004$); there was no difference between *muss p* and *wohl p* (β=-.02, *SE*=0.03, *t*=-.61, $p< .55$). This shows, similar to what we found in the domain of production, that perceived speaker commitment in the case of both epistemic *must* and discourse particles is stronger than in the case of using otherwise synonymous adverbs such as *vermutlich*.

These results mirror the effects found in Experiment 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

[7] This experiment can be viewed here. The German version can be viewed here.

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, both in English (β=.14, SE=0.03, t=4.7, p<0.0001) and in German (β=.08, SE=0.04, t=2.23, p< .03). Note that this finding suggests that *must* is not special in generating the effect of stronger speaker commitment than resulting listener belief.

As in Experiment 3a, to address whether inferred speaker evidence strength mirrors production, we conducted another mixed effects linear regression, predicting inferred strength of evidence for *p* from a dummy-coded utterance predictor with *must/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: again, participants inferred stronger evidence was available to the speaker after observing the bare utterance than *must/muss p* in both English (β=.06, SE=.02, t=3.2, p< .002) and German (β=.08, SE=.02, t=4.05, p< .0001). However, inferred evidence strength was no different in English for *probably p* (β=-.02, SE=0.02, t=-1.11, p< .27) or *might p* (β=-.01, SE=0.02, t=-.49, p< .63); nor in German for *vermutlich p* (β=-.0003, SE=.02, t=-.02, p< .99) or *wohl p* (β=.008, SE=0.02, t=.36, p< .72).

Allowing this model to interact with a belief holder predictor and applying it simultaneously to the Experiment 3a dataset yields no main effect of belief holder in either English (β=.03, SE=.02, t=1.53, p< .13) or German (β=.0008, SE=0.03, t=.04, p< .97). This finding is unsurprising and serves as a sanity check for the belief holder effect, given that this aspect of the dependent measure was identical across experiments.

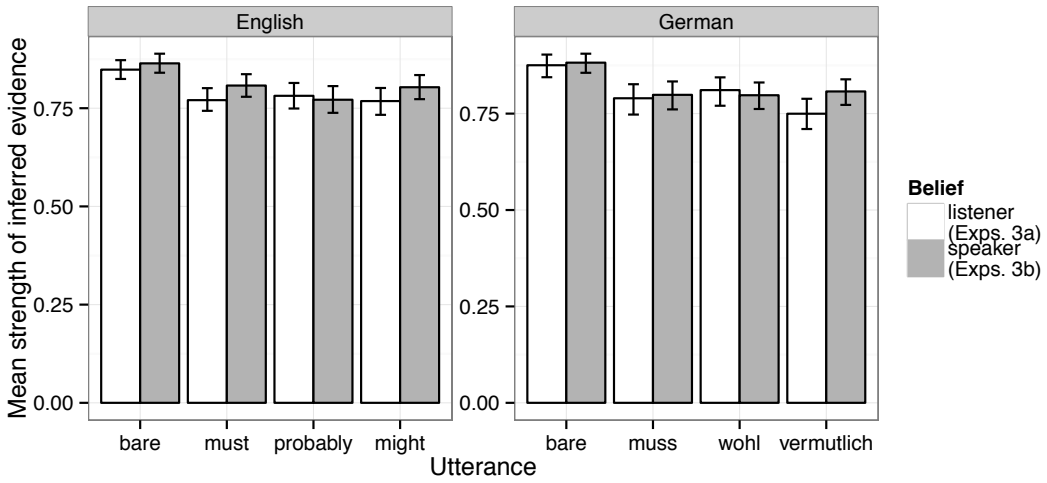

Figure 6: Mean inferred evidence strength by utterance for English (left) and German (right). Error bars indicate 95% bootstrapped confidence intervals.

6 General conclusion and outlook

In this paper, we presented a series of experiments focusing on the extent to which listeners' interpretation of certain types of evidential devices and their judgments about speaker commitment differ in strength. We demonstrated that speakers' production preferences for these different devices mirror the comprehension results under varying evidential

circumstances. By investigating the use of evidential devices from this perspective, we introduced a new experimental paradigm for exploring the impact of different evidential circumstances on the use of modal evidentials, epistemic discourse particles, and statements with no evidential markers at all. In other words, we introduced an experimental paradigm for mapping the empirical terrain of evidential devices.

Although our experimental data cannot decide fully between the various theoretical accounts on challenging phenomena such as epistemic *must* (Section 1.1) and discourse particles (Section 1.2), resolving these debates requires systematically testing the compatibility of these evidential devices with different degrees of evidence strength and evidential circumstances; we have presented an experimental paradigm for doing just that. This paradigm will allow for the exploration of how different evidential devices are used and comprehended compared to alternative lexical choices. It also highlights the central role of speaker commitment in theories of evidential devices. Below, we comment on some of the conclusions we can draw from our results.

As for discourse particles, no study has to date taken into account the component of different degrees of speaker commitment when debating how, for instance, closely related sentence adverbs can be distinguished from synonymous discourse particles at the level of semantics. This is a new data point that needs to be accounted for in our theoretical understanding of the lexical inventory of evidential words in the German language.

As for epistemic *must*, we could, as expected, confirm the claim that the epistemic use of *must* expresses a weaker commitment than the bare form. However, in addition to experimentally confirming this (admittedly) trivial observation, we have tested for different degrees of evidence strength and also compared epistemic *must* not only to the bare form but also to alternative modal expressions available in the English lexicon (i.e., adverbs and *might*). In this context, we found that *must* is used in evidential circumstances where speaker commitment can be considered rather strong. This appears to concur with claims in the literature that highlight this strong component of *must* (e.g., von Fintel & Gillies 2010, 2016). However, the results also suggested that speaker commitment is lower for *must* than for the bare form. This is at odds with von Fintel & Gillies' claim that "[s]peakers who say *must* Φ are just as strongly committed to the prejacent as those who assert Φ by itself" (von Fintel & Gillies 2010: 30). This result might also serve to explain concrete felicity effects in the lexical domain of modal expressions as recently described by Matthewson & Truckenbrodt (2017). They point out that epistemic *must*/*muss* is in sharp contrast to the epistemic use of other modals such as *sollen* in German. In particular, *must*/*muss* is considerably stronger in the domain of speaker commitment. Consider one of their German examples (Matthewson & Truckenbrodt 2017: 12):

(12) [Heike says that Maria is in the kitchen, but I am not convinced, since I think I would have seen her go into the kitchen. I say:]
   a. # Maria **muss** in der Küche sein (aber ich habe meine Zweifel).
      Maria must in the kitchen be (but I have my doubts) 'Maria must be in the kitchen (but I have my doubts).'
   b. Maria **soll** in der Küche sein (aber ich habe meine Zweifel).
      Maria SOLL in the kitchen be (but I have my doubts) 'Maria is supposed to be in the kitchen (but I have my doubts).'

The context in (12) provides a report in the context, and epistemic *muss*, in contrast to *sollen*, requires a strong commitment to the prejacent and can thus not be used in a context where the speaker might have doubts to a certain extent. We hypothesize that both *probably* and *might* (our tested items) pattern with *soll* rather than with *muss*/*must*, meaning that

muss/*must* conveys a rather high degree of speaker commitment, and this is what we found in our experiments.

All in all, given that the theoretical literature both on English and on German evidential words is often based on subtle judgments of utterances, our paper presents for the first time an experimental investigation on cross-linguistic expressions conveying different strengths of speaker commitment. Our new experimental paradigm thus illustrates a new approach to detect and define meaning differences in the lexical inventory of evidentials across languages and hence provides a good starting point for approaching theoretical debates on the nature of evidential expressions from a fresh, experimentally-driven perspective.

Acknowledgments

TO BE INSERTED UPON REVIEW

References

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford: Oxford University Press.

Cardinaletti, A. (2011). German and Italian modal particles and clause structure. *The Linguistic Review* 28, 493–531.

Döring, S. and S. Repp (2016). The modal particles *ja* and *doch* and their interaction with discourse structure: Corpus and experimental evidence. To appear in S. Featherston, R. Hörnig, S. von Wietersheim & S. Winkler (Eds.), *Information Structure and Semantic Processing*. Berlin: de Gruyter.

Dörre, L. and A. Trotzke (2017). The processing of secondary meaning: An experimental comparison of focus and modal particles in *wh*-questions. To appear in D. Gutzmann and K. Turgay (Eds.), *Secondary Meaning and Linguistic Encoding*. Leiden: Brill.

Ernst, T. (2007). On the role of semantics in a theory of adverb syntax. *Lingua* 117, 1008–1033.

Giannakidou, A. and A. Mari (2016). Epistemic future and epistemic MUST: Nonveridicality, evidence, and partial knowledge. In J. Blaszczak, A. Giannakidou, D. Klimek-Jankowska, and K. Migdalski (Eds.), *Mood, Aspect, Modality Revisited: New Answers to Old Questions*, pp. 75–124. Chicago: The University of Chicago Press.

Goodhue, D. (2016). Epistemic *must* is not evidential, it's epistemic. *Proceeding of NELS* 46, 321–334.

Grosz, P. (2016). Discourse particles. To appear in L. Matthewson, C. Meier, H. Rullmann, and T. E. Zimmermann (Eds.), *The Companion to Semantics (SemCom)*. Oxford: Wiley.

Grosz, P. (2017). Shedding new light on the *wohl* muddle: The particle *schier* in Austrian German. *Wiener Linguistische Gazette* 82, 71–78.

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

Knobe, J. and S. Yalcin (2014). Epistemic modals and context: Experimental data. *Semantics and Pragmatics* 7, 1–21.
Korotkova, N. (2016). Heterogeneity and uniformity in the evidential domain. PhD dissertation, University of California, Los Angeles.

Kratzer, A. (1991). Modality. In A. von Stechow and D. Wunderlich (Eds.), *Semantics: An International Handbook of Contemporary Research*, pp. 639–650. Berlin: de Gruyter.

Lassiter, D. (2014). The weakness of *must*: In defense of a mantra. *Semantics and Linguistic Theory* 24, 597–618.

Lassiter, D. (2016). *Must*, knowledge, and (in)directness. *Natural Language Semantics* 24, 117–163.

Mandelkern, M. (2017). *Coordination in Conversation*. PhD dissertation, MIT.

Matthewson, L. (2012). Evidence about evidentials: Where fieldwork meets theory. In B. Stolterfoht and S. Featherston (Eds.), *Empirical Approaches to Linguistic Theory: Studies of Meaning and Structure*, pp. 85–114. Berlin: de Gruyter.

Matthewson, L. (2015). Evidential restrictions on epistemic modals. In L. Alonso-Ovalle and P. Menéndez-Benito (Eds.), *Epistemic Indefinites. Exploring Modality Beyond the Verbal Domain*, pp. 141–160. Oxford: Oxford University Press.

Matthewson, L. and H. Truckenbrodt (2017). Modal flavour/modal force interactions in German: *soll*, *sollte*, *muss* and *müsste*. Ms.

Murray, S. (2017). *The Semantics of Evidentials*. Oxford: Oxford University Press.

Salmon, W. (2011). Conventional implicature, presupposition, and the meaning of *must*. *Journal of Pragmatics* 43, 3416–3430.

Ünal, E. and A. Papafragou (2018). Evidentials, information sources, and cognition. In A. Y. Aikhenvald (Ed.), *The Oxford Handbook of Evidentiality*, pp. 175–184. Oxford: Oxford University Press.

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

von Fintel, K. and A. S. Gillies (2016). Still going strong: The semantics and pragmatics of epistemic *must*. Ms.

Zimmermann, M. (2004). Zum *wohl*: Diskurspartikeln als Satztypmodifikatoren. *Linguistische Berichte* 199, 253–286.

Zimmermann, M. (2008). Discourse particles in the left periphery. In B. Shaer, P. Cook, W. Frey, and C. Maienborn (Eds.), *Dislocated Elements in Discourse*: *Syntactic, Semantic, and Pragmatic Perspectives*, pp. 200–231. New York & London: Routledge.

A Pieces of evidence

This section lists, for each proposition *p*, 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 können hören, 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 würde.

4. You see a person come in from outside with wet hair and wet clothes.
   Sie sehen, wie jemand mit nassen Haaren und durchnässten Kleidern von draußen 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 berühren 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 hören 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 Kopfhörer auf und hören 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 Kopfhörer auf und hören Musik, sehen aber aus dem Fenster und beobachten, wie der Postbote vor der Nachbarstür einen Satz nach hinten macht.

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