Evidential uncertainty involves both pragmatic and extralinguistic reasoning: a computational account

Alon Fishman
Department of Linguistics, Tel Aviv University
Tel Aviv, Israel 69978

Judith Degen
Department of Linguistics, Stanford University
Stanford, CA USA 94305

Abstract

Using evidential expressions to indicate one's source of information for an utterance tends to convey uncertainty on the speaker's part. Previous accounts of this uncertainty inference attribute it to either extralinguistic reasoning about evidence directness, or to pragmatic reasoning about alternative utterances. Here we present a novel hybrid account, and introduce a set of utterances which allows us to tease apart the three accounts' predictions. We test these predictions in two studies by manipulating the directness of evidence indicated by an evidential expression. Exp. 1 shows that listeners infer more uncertainty with extreme values of directness. Exp. 2 shows that speakers are more likely to indicate evidence in contexts where the evidence is unreliable. We argue that these findings support an account which involves both extralinguistic and pragmatic reasoning, and develop a formal implementation of such an account within the Rational Speech Act framework.

Keywords: experimental pragmatics; probabilistic pragmatics; psycholinguistics; evidentials

Introduction

One of the principal uses of language is to exchange information about the world, but this information can be uncertain. Speakers might only have indirect or unreliable evidence for a belief, and a cooperative speaker may seek to convey this. Speakers can make their evidence explicit using various *evidential devices*, including lexical items, embedding constructions, and in some languages, an obligatory grammatical category (Aikhenvald, 2014; Murray, 2021). For instance, *the dress* ***looks*** *new* implies that the dress' appearance evidences that it is new (Gisborne, 2010; Muñoz, 2019), ***I hear*** *the dress is new* implies a secondhand report (Simons, 2007), and *the dress* ***must*** *be new* implies an indirect inference (Mandelkern, 2019; von Fintel & Gillies, 2010).

Using (non-obligatory) evidential devices is typically judged to convey uncertainty regarding the evidenced proposition, relative to using an unmarked bare form, e.g., *the dress is new* (Pogue & Tanenhaus, 2018). This uncertainty is not part of the utterance's semantic content but rather a defeasible inference which is affected by context and prosody (Kurumada et al., 2014; Kurumada et al., 2012; Speas, 2018).

What causes the uncertainty inference? One approach, alluded to by von Fintel and Gillies (2010) and Mandelkern (2019), would attribute it to extralinguistic reasoning about evidence type, specifically about evidence directness. Presumably, for any given proposition, some types of evidence offer more direct support than others (Matthewson, 2020, regarding the relation between evidence type and directness). On this approach, listeners ascribe to the speaker a degree of certainty proportional to the directness of evidence available to the speaker, as implied by the evidential device they used.

A second approach, advanced by Degen et al. (2015), attributes the uncertainty inference to Gricean reasoning, that is, to pragmatic reasoning about alternative utterances the speaker could have used, but chose not to. In particular, the speaker could have used a bare form with no evidential device, which by default would have conveyed maximal certainty. Listeners reason that the speaker must have had a communicative intent in using the more costly marked form, and that intent could be to avoid conveying the same certainty.

These two approaches make conflicting predictions regarding a set of utterances largely overlooked by researchers, ones in which an evidential device implies the most direct type of evidence possible for a proposition. Consider *the dress looks blue*, which implies that the dress' appearance evidences that it is blue. On the face of it, this is the most direct type of evidence possible to support a belief about an object's color. If the speaker is ascribed a degree of certainty proportional to the directness of their evidence – via extralinguistic reasoning – then the uncertainty inference should be substantially abated. In contrast, if the marked utterance is compared to the less costly bare form alternative – via Gricean reasoning – then the uncertainty inference should arise as usual.

Further complicating the picture is an observation by Grice (1961) regarding utterances which imply direct evidence, namely that they are only natural in contexts where there is reason to doubt or deny the evidenced proposition. For instance, *the dress looks blue* is odd when observing the dress in broad daylight, but natural when examining a famously contentious picture of the dress on a screen. This 'doubt-or-denial' condition is purportedly an extra discourse effect, beyond the ordinary uncertainty inference associated with evidential devices. Moreover, it raises the possibility that utterances implying direct evidence would convey even more uncertainty than ones implying indirect evidence.

We propose a hybrid account of the discourse effects of evidential devices, which relies on both Gricean reasoning and extralinguistic reasoning about evidence type. On our proposal, all utterances with evidential devices are compared with unmarked bare form alternatives. This routinely triggers an uncertainty inference via Gricean reasoning, regardless of the type of evidence implied. For indirect evidence, this aligns with listeners' extralinguistic assumptions, since world knowledge dictates that indirect evidence coincides with low speaker certainty. But for direct evidence, an additional step of extralinguistic reasoning is required, because direct evidence is expected to coincide with high certainty. To solve this clash in expectations, listeners reason that the direct evidence in these particular circumstances must be compromised, thereby deriving Grice's doubt-or-denial condition.

Our proposal makes two concrete predictions about utterances implying direct evidence: (i) they convey uncertainty relative to bare forms, and (ii) they are more likely to be used under circumstances where the implied type of evidence is compromised and hence less reliable, e.g., poor visibility in the case of visual evidence. We present results from two experiments, based on the experimental paradigm of Degen et al. (2019), which provide initial corroboration of our predictions. Exp. 1 measures the perceived certainty of a speaker using bare forms and using evidential devices implying evidence of varying directness. Exp. 2 explores the choice between using bare forms and using evidential devices implying evidence of varying directness, in contexts with either good or poor perceptibility conditions.

In both experiments, evidence directness is based on perceptual strength norms collected by Lynott and Connell (2009). Lynott and Connell generated perceptual strength norms for the five classical senses for 423 adjectives, by asking participants how strongly a property was experienced by seeing, hearing, feeling through touch, etc. We use an adjective's perceptual strength for a given sense as the directness of evidence obtained through that sense for a proposition about the adjective. To illustrate, "blue" has higher visual strength than "new", hence visual evidence is more direct for the proposition *the dress is blue* than for the proposition *the dress is new*. Consequently, *the dress* ***looks*** *blue* implies more direct evidence than *the dress* ***looks*** *new*.

Finally, we develop a computational model of our hybrid account, implemented in the Rational Speech Act (RSA) framework (Frank & Goodman, 2012), and show that it can derive some of the qualitative findings of the experiments. Our model extends the basic RSA framework with formal representations of the directness and reliability of the evidence available to the speaker. These are taken as input by the speaker's belief function, and are included in the set of inferences outputted by the pragmatic listener function.

Experiment 1: interpretation

We investigate how the directness of evidence available to the speaker, as implied by use of an evidential device, affects the speaker's perceived certainty. Participants are placed in a listener's role: they are presented with utterances and asked to rate the speaker's certainty.[1]

Degen et al. (2019) previously established that listeners ascribe varying degrees of certainty to a speaker depending on whether they use an evidential device, and on which evidential device they use. Importantly, all the evidential devices they examined (English *must*, *might*, and *probably*, and German *muss* ('must'), *vermutlich* ('probably') and *wohl* (lit. 'well')) imply the speaker has indirect evidence. As such, their results could be attributed to either extralinguistic reasoning about evidence directness, or to Gricean reasoning about competition with the bare form. By manipulating the directness of evidence, the present experiment can tease these two accounts apart. In addition to utterances with evidential devices, bare utterances are included as a baseline condition.

Method

Participants. We recruited 40 participants on Prolific.

Materials and procedure. Participants read the following story introducing the speaker and the discourse context:

Your friend Taylor is at a party which you could not attend. The party is pretty fancy, but also crowded and noisy. Over the course of the party, Taylor is texting you about the people at the party and what they are wearing. Each of the following statements is a text you receive from Taylor.

The speaker was described as the listener's friend to alleviate "epistemic vigilance" targeted at the risk of being intentionally misinformed (Sperber et al., 2010). The party was described as crowded and noisy to facilitate the assumption that the speaker may be uncertain about what they observe.

Participants then read a total of 20 statements, each describing a person's item of clothing using a single adjective. Critical items were 5 bare utterances with no evidential device, and 5 utterances with the perception verb *looks*. There were also 10 filler items, 5 with the modal *might* and 5 with the phrase *I think*. Examples are given below:

Elliot's suit is beige (bare)
Holly's dress looks purple (*looks*)
Mark's outfit might be crimson (*might* filler)
I think Tom's vest is green (*think* filler)

After each statement, participants were asked about the speaker's certainty regarding the relevant proposition, e.g., "Is Taylor sure that Elliot's suit is beige?" and adjusted a slider with endpoints labeled "Absolutely sure" (coded as 1) and "Not sure at all" (coded as 0).

After the first statement they saw, and after 4 other statements, participants were additionally asked about the speaker's evidence for the proposition, e.g., *Why does Taylor think that Elliot's suit is beige?* and chose between four potential sources of evidence: visual (*Taylor saw it*), haptic (*Taylor touched it*), reported (*Someone told Taylor about it*), or indirect inference (*Taylor has known Elliot for a long time*). These questions were included to ensure that participants paid attention to the speaker's choice of evidential device.

Adjectives on critical trials were randomly selected from the list of adjectives examined by Lynott and Connell (2009). We used only adjectives which were familiar to all of their participants, had a frequency greater than 1 in the British Na-

[1] Experimental materials, data, and analysis scripts are available at https://github.com/AlonFishm/Evidential_uncertainty.

tional Corpus (BNC), and were not predicates of personal taste or aesthetic judgment, as those allow a non-evidential reading of *looks* (McNally & Stojanovic, 2017; Poortvliet, 2018). The 10 adjectives used in Exp. 1 and their mean visual strength ratings (on a scale of 0 to 5) were: *purple* (5.00), *shiny* (4.95), *short* (4.95), *clean* (4.62), *striped* (4.52), *beige* (4.48), *bulky* (4.43), *loose* (4.14), *oily* (3.90), and *fuzzy* (3.67).

Two lists were created, each containing 5 adjectives in bare utterances and 5 with the evidential device *looks*, along with the 10 filler items. Participants were randomly assigned to a list, and items were presented to them in pseudo-random order (a condition did not occur twice in a row).

Results and discussion

Mean certainty ratings are shown in Fig. 1. To assess the effects of perceptual strength and utterance type on interpretation, we conducted a mixed-effects regression predicting speaker certainty from centered fixed effects of utterance type (reference level before centering: “bare”) and visual strength, a second-order polynomial term for the visual strength predictor,[2] the two-way interactions of utterance type and visual strength, as well as by-participant and by-adjective random intercepts (the most complex random effects structure that allowed the model to converge).

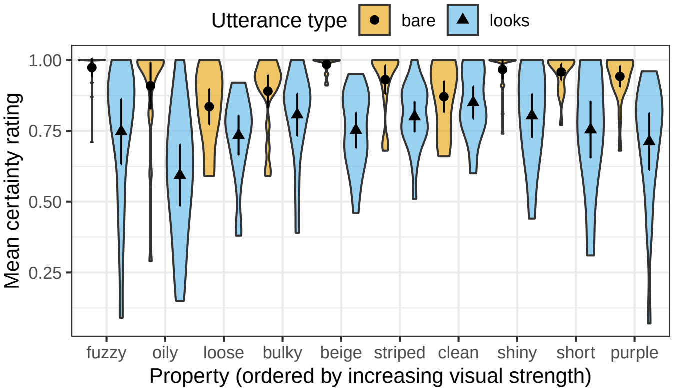

Figure 1: Mean certainty ratings in Exp. 1 by adjective and utterance type. Error bars indicate 95% bootstrapped confidence intervals, violin plot shows data distribution.

Speaker certainty was rated lower for utterances with the evidential device than for bare utterances ($\beta = -0.14, SE = 0.02, t = -6.96, p < .0001$). There was no main effect of visual strength ($\beta = 0.04, SE = 0.03, t = 1.32, p < .23$) or squared visual strength ($\beta = 0.01, SE = 0.07, t = 0.16, p < .16$), but the interaction between utterance type and squared visual strength was significant: for utterances with the evidential device, certainty was lower with extreme values of visual strength ($\beta = -0.16, SE = 0.08, t = -1.98, p < .05$).

The results of the current experiment support the predictions of the Gricean account, in that the use of an evidential device always conveys uncertainty relative to a bare utterance. The results could also be interpreted as supporting the predictions of an extralinguistic account: the uncertainty effect of using an evidential device depends on the directness of the evidence implied. However, contrary to the prediction that speaker certainty is proportional to evidence directness, the relation between them appears to be quadratic rather than linear. Specifically, the uncertainty effect is amplified both for relatively indirect evidence and for maximally direct evidence, the latter of which cannot be explained by a fully extralinguistic account.

We take these results as supporting a hybrid account that involves both Gricean reasoning and extralinguistic reasoning about evidence type. We predict that the increased uncertainty associated with utterances implying maximally direct evidence is the result of an extra inference, that the evidence is compromised, for instance by poor perceptibility conditions. This prediction is tested in Exp. 2.

Experiment 2: production

We investigate how evidence directness and perceptibility conditions affect the choice between using a bare utterance and using an evidential device. Participants are placed in a speaker’s role: they are presented with a context and asked to choose between possible utterances.

Pogue and Tanenhaus (2018) and Degen et al. (2019) have both explored choice of utterance as a function of the evidence available to the speaker. Pogue and Tanenhaus presented participants with visual evidence in the form of images, and manipulated the completeness of the images and the amount of time participants had to view them. Degen et al. presented participants with textual descriptions of various types of evidence: perceptual, reported, and inferential.

Essentially, Pogue and Tanenhaus kept the type of evidence constant and manipulated its reliability, while Degen et al. did the reverse, manipulating evidence type and not reliability. Hence in both studies, the evidence presented to participants could be ranked on a single scale of evidence strength. Both studies found that speakers were more likely to use an evidential device with weak evidence, and more likely to use a bare utterance with strong evidence.

The present experiment explores a more complex notion of evidence strength, comprising two elements: directness and reliability. As in Exp. 1, evidence directness is based on Lynott and Connell (2009)’s perceptual strength norms for adjectives. Unlike Exp. 1, three sensory modalities are included: visual, auditory, and haptic. Evidence reliability is manipulated with textual descriptions of perceptibility conditions.

Method

Participants. We recruited 40 participants on Prolific. Due to a technical issue, the responses of 7 participants were lost, leaving us with responses from 33 participants.

Materials and procedure. Participants saw 12 texts describing situations and chose an utterance to produce in each situation. Each text described the speaker standing outside a

[2] Inclusion of this term was motivated by the consideration that decreased certainty on the lower end of the scale is predicted by extralinguistic reasoning and on the upper end of the scale by Grice (1961)’s doubt-or-denial condition.

room, about which they have only a single source of information. Participants were asked what they would say to a friend without access to the same information, if the friend asked them whether the room had a certain property.

Critical items were 6 situations in which the participants' source of information was visual (a window they could look through), auditory (a door they could listen at), or tactile (a gap under the door they could reach through). There were also 6 filler items in which the participants' source of information was olfactory, linguistic, or mixed. In each situation, the source of information was described as either good or poor. Examples of situations with good visual evidence and poor auditory evidence, respectively, are the following:

Imagine that you are standing outside a room. You can't hear anything inside, but there is a window that you can look through. The window is perfectly clear, so you can see what it's like in the room very well.

Imagine that you are standing outside a room. You can't see inside, but you can listen at the door. However, the door is very thick, so it's difficult to hear what it's like in the room.

On critical trials, participants were asked to choose between a bare utterance and an utterance with a perception verb matching the source of information: *looks*, *sounds*, or *feels*. On filler trials, participants were asked to choose between two utterances with two different perception verbs.

Adjectives on critical trials were selected from the list of adjectives examined by Lynott and Connell (2009). We used only adjectives which were familiar to all of their participants, had a frequency greater than 1 in the BNC, and were not predicates of personal taste or aesthetic judgment. We selected "weather" predicates, which could occur in impersonal constructions, e.g., *it's hot in there*. The 6 adjectives used in Exp. 2 and their mean perceptual strength ratings for the three studied modalities (on a scale of 0 to 5) are given in Table 1.

Table 1: Adjectives used in Exp. 2.

Adjective	Visual	Auditory	Haptic
bright	5.00	0.14	0.19
crowded	4.62	3.71	2.29
wet	4.33	1.86	4.67
hot	3.33	1.05	4.86
humid	1.76	0.24	3.29
noisy	1.67	4.95	0.29

Six lists were created, each containing two situations per information source, one good and one poor. Participants were randomly assigned to a list, and items were presented to them in pseudo-random order (a condition did not occur twice in a row). The order in which utterances were presented as choices was alternated between critical items.

Results and discussion

Proportions of evidential device use are shown in Fig. 2. We conducted a mixed-effects logistic regression predicting the log odds of using an evidential device from a dummy coded fixed effect of perceptibility (reference level: "good"), a centered fixed effect of perceptual strength, a second-order polynomial term for the perceptual strength predictor, the two-way interactions between perceptibility and perceptual strength, as well as by-participant and by-adjective random intercepts (the most complex random effects structure that allowed the model to converge).

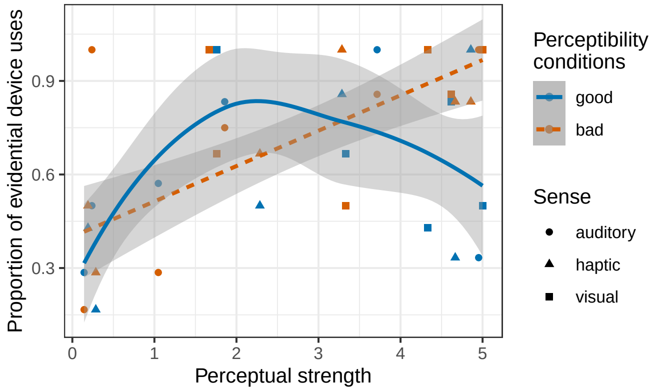

Figure 2: Proportion of participants choosing to use an evidential device (and not a bare utterance) in Exp. 2, by perceptual strength, perceptibility conditions, and sensory modality. Regression lines estimated for good and poor perceptibility, shading indicates 95% confidence region.

There was no main effect of perceptibility conditions ($\beta = -0.45, SE = 0.61, p < .47$) or perceptual strength ($\beta = 0.14, SE = 0.13, p < .27$). However, there was a main effect of squared perceptual strength ($\beta = -0.32, SE = 0.11, p < .01$), such that participants were less likely to use evidential devices with extremely low or high perceptual strength when perceptibility conditions were good. Both interaction terms reached significance, suggesting both that in poor perceptibility conditions the likelihood of using an evidential device increased linearly with perceptual strength ($\beta = 0.48, SE = 0.20, p < .05$) and that, different from the "good" condition, there was no endpoint effect ($\beta = 0.29, SE = 0.16, p < .07$).[3]

To compare the results of Exp. 2 to those obtained in Exp. 1, we need to reconstruct the notion of evidence strength from its two material components: directness and reliability. Evidence is strongest when it is both direct, represented here by high perceptual strength, and reliable, represented here by good perceptibility conditions. With strong evidence, we expect high speaker certainty, and consequently more bare utterances. This prediction is borne out. As evidence becomes weaker, due to either indirectness or unreliability, we expect more uncertainty, and consequently more evidential devices. This prediction is borne out as well.

Where our predictions seem to fail is with the very weakest evidence, where we would expect the greatest degree of uncertainty, but find minimal use of evidential devices. We attribute this to a limitation of the experimental design: participants were forced to choose between a bare utterance and an utterance with an evidential device, in a context where neither option is apt. In general, answering a question based on

[3] A model that included a control predictor for sensory modality replicated the results and yielded no significant effects of modality.

very weak evidence (e.g., answering whether a room is noisy based only on haptic evidence) is not an exemplar of cooperative pragmatic behavior. Many speakers in such circumstances would prefer to admit ignorance or say nothing, but these options were not available to our participants.

We can think of two reasons why our participants were particularly averse to using an evidential device when the evidence implied by it was very weak. First, it's possible that a minimum threshold of evidence strength is "hardcoded" into the semantics of evidential devices. In other words, utterances such as *it feels noisy in there* may be grammatically unacceptable, in addition to pragmatically uncooperative. Conversely, any constraint on the evidence supporting bare utterances is solely pragmatic (e.g., the Gricean maxim of quality). A second possibility is that a minimum of evidence strength is required for an utterance with an evidential device to be interpreted as addressing the question under discussion (Roberts, 2004), otherwise it comes off as a non sequitur. Again, this is not a pitfall bare utterances can fall into. These two tentative explanations could potentially be teased apart experimentally, but we leave this to future research.

Formal model

We formalize our account within the Rational Speech Act (RSA) framework (Frank & Goodman, 2012). The framework models cooperative communication between rational agents using recursive Bayesian inference. At the base of the recursion sits a "literal listener", L_0, a function from an utterance to a probability distribution over states in which the utterance is literally true. Next, the "pragmatic speaker", S_1, chooses an utterance, seeking to maximize informativity (probability of correctly inferring the intended meaning) to the literal listener while minimizing utterance cost. The pragmatic speaker is typically represented as a function from an observed state (intended meaning) to a probability distribution over utterances. Finally, the "pragmatic listener," L_1, uses Bayes' rule and inverts the pragmatic speaker's utterance probabilities, re-weighted by a prior on states, to recover the most likely state intended by the speaker.

In our model, the space of possible utterances includes bare utterances, utterances with each of the evidential devices *looks*, *sounds* and *feels*, and a "null utterance" (saying nothing). For any at-issue proposition q, the utterance space is:

$$U_q = \{q, looks(q), sounds(q), feels(q), \text{NULL}\} \qquad (1)$$

We assume simplified semantics for these utterances: bare q is true iff q holds; for any evidential device EVID that implies evidence of type ε, $\text{EVID}(q)$ is true iff there is evidence of type ε that q holds; and NULL is always true.[4] To capture that evidence does not always attest to actual fact, and that different types of evidence may be at odds, we represent states of the world as n-tuples. The first element in the n-tuple is the "actual" state of the world, to which bare utterances refer. Subsequent elements are "evidence states", each of which represents whether or not there is evidence of a particular type for the actual state. Thus, for any proposition q and evidence for q of types $\varepsilon \dots \varepsilon'$, the set of possible states is

$$S_q = \{q, \neg q\} \times \{\varepsilon_q, \neg\varepsilon_q\} \times \dots \times \{\varepsilon'_q, \neg\varepsilon'_q\} \qquad (2)$$

The above representation captures the fact that evidence does not always match reality. However, we also want to capture the intuition that evidence does tend to match reality more often than not, otherwise it wouldn't count as evidence. We incorporate this intuition into listeners' prior beliefs about the state of the world, using the parameter $E > 1$, the "evidence coefficient":

$$P((q, \varepsilon_q)) = P((q, \neg\varepsilon_q)) \cdot E \qquad (3)$$

This means that the prior probability of q is greater by a factor of E when there is evidence matching it compared to when there isn't. This is what allows utterances with evidential devices to address questions about the actual state of the world. It ensures that the literal listener function, upon observing an utterance with an evidential device, will assign a higher probability to the evidenced proposition than to its negation (we otherwise assume flat priors over states). For an utterance u addressing proposition q, L_0 sums the prior probabilities of states $s_{u,q}$ where both u and q are true to compute the probability of q:

$$P_{L_0}(q|u) \propto \sum P(s_{u,q}) \qquad (4)$$

Given that speakers may have uncertainty about the true s, we represent speaker beliefs as a probability distribution over possible states (Goodman & Stuhlmüller, 2013; Scontras et al., 2018, for a similar approach). Here, we assume that speakers' beliefs are based on the DIRectness and RELiability of the evidence available to them. DIRectness is a function of evidence type and proposition, so that for instance, visual evidence is more direct than haptic evidence for *it's bright in there*, but less direct for *it's hot in there*. For any proposition q and evidence for q of type ε: $\text{DIR}(\varepsilon_q) \in (0,1)$. RELiability is in principle a function of evidence type and context, so that for instance, dim lights make visual but not haptic evidence less reliable. For simplicity, we assume just two possible values for reliability, corresponding to the good and poor perceptibility conditions in Exp. 2: $\text{REL}(good), \text{REL}(poor) \in (0,1)$. Thus, given evidence for proposition q of type ε in conditions v, the speaker's belief regarding q is

$$P_{S_1}(q|\varepsilon_q, v) = 0.5 + \frac{\text{DIR}(\varepsilon_q) \cdot \text{REL}(v)}{2} \qquad (5)$$

With maximally direct and reliable evidence, the speaker's belief in the proposition approaches absolute certainty (1), but when evidence is either very indirect or very unreliable, belief approaches agnosticism (0.5).

In standard RSA models, the speaker does not produce utterances whose truth they are not certain of. This requirement is too strong for our purposes, as it predicts that bare utterances would require perfect certainty about the actual state of the world, and hence would essentially never be used. We

[4] The semantics of evidentials, as well as the logical and pragmatic relations between evidentials and the propositions they evidence, are much debated topics (Faller, 2020; Korotkova, 2020; McCready, 2020; Murray, 2020, for recent proposals and discussion).

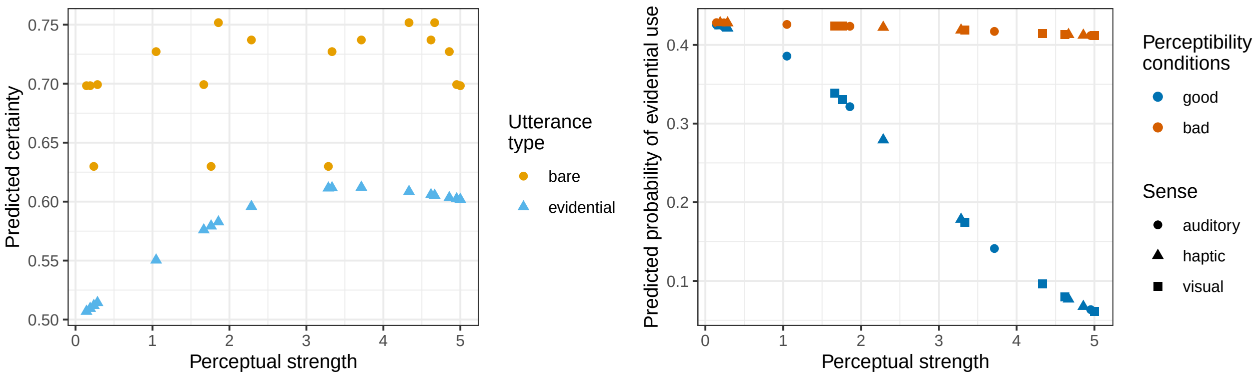

Figure 3: Model predictions. Left: pragmatic listener's inferred speaker beliefs. Right: pragmatic speaker's evidential use.

therefore adopt the notion of relaxed semantics from Degen et al. (2020). The idea is that informativity is computed with respect to a non-deterministic continuous semantics instead of a deterministic Boolean semantics. This is captured by the literal listener ascribing a small but non-zero probability to states in which the input utterance is not literally true.

The pragmatic speaker function, in choosing an utterance, seeks to minimize the (Kullback-Leibler) divergence between their own belief state $P_{S_1}(\cdot|\varepsilon_q, v)$, and the expected belief state of the literal listener $P_{L_0}(\cdot|u)$. The parameter $\alpha > 0$ represents the speaker's "optimality", which is to say, how committed they are to choosing the most informative utterance:

$$P_{S_1}(u|\varepsilon_q, v) \propto exp(-D_{KL}(P_{S_1}(\cdot|\varepsilon_q, v)||P_{L_0}(\cdot|u)) \cdot \alpha) \qquad (6)$$

The last element in our model is the pragmatic listener function. The pragmatic listener performs joint inference over the true state and the pragmatic speaker's evidential state. Our pragmatic listener outputs probability distributions over actual states, type of evidence available to the speaker, and conditions under which the evidence was perceived. It does so by using Bayes' rule to invert the pragmatic speaker's utterance probabilities:

$$P_{L_1}(q, \varepsilon_q, v|u) \propto P_{S_1}(u|\varepsilon_q, v) \cdot \sum P(s_{u,q}) \qquad (7)$$

We use the model to simulate predictions for Exps. 1 and 2. For the evidence coefficient E, we use the average certainty rating obtained in Exp. 1 for utterances with evidential devices, divided by agnosticism: $E = \frac{0.75}{0.5} = 1.5$. We generate directness values by dividing Lynott and Connell (2009)'s perceptual strength norms by 5,and set $\alpha = 10$, $\mathrm{REL}(good) = 0.9$ and $\mathrm{REL}(poor) = 0.1$. The model's predictions with these parameter values are shown in Fig. 3.

We engage only in qualitative model comparison, since the collected data are not suited for quantitative evaluation, due to limitations of the experimental designs: Exp. 1 included only a single sensory modality and a relatively narrow range of perceptual strength ratings; Exp. 2 forced participants to choose between two potentially inappropriate utterances, as discussed above. Nevertheless, the model is able to capture some of the qualitative findings of the experiments. For Exp. 1, the pragmatic listener function consistently associates higher certainty with the bare utterance than with the evidential device. Moreover, given utterances with an evidential device as input, it produces an n-shaped quadratic relation between perceptual strength and certainty: certainty decreases with extreme values of perceptual strength. For Exp. 2, the pragmatic speaker's probability of using an evidential device decreases with perceptual strength in good perceptibility conditions, but not in poor perceptibility conditions. In other words, the model is able to derive both the general uncertainty inference associated with evidential devices, and the more specific doubt-or-denial condition on utterances implying maximally direct evidence. The next, more rigorous test of the model will involve extended designs of Exps. 1 and 2: the former covering more sensory modalities and a wider range of perceptual strength ratings, and the latter offering additional utterance options to the participants.

Conclusion

We have shown that the uncertainty inference associated with evidential devices is more complex than predicted by previous accounts. Exp. 1 revealed that the use of an evidential device conveyed uncertainty even when implying maximally direct evidence. Moreover, the uncertainty inference was enhanced for maximally direct evidence. These findings challenge extralinguistic accounts which predict that perceived speaker certainty is directly proportional to the directness of evidence available to the speaker, as well as purely Gricean accounts which predict no effect for evidence directness.

We believe that only a hybrid account can explain the results of Exp. 1. Our account does so by introducing a notion of evidence strength comprised of two distinct elements: directness and reliability. The account's predictions were partly borne out in Exp. 2, which revealed that evidential devices were most likely to be used with either high directness and low reliability, or medium directness and high reliability.

Our account could be developed further to make predictions for languages with a grammatical category of evidentiality, in which information source is obligatorily marked on every sentence. If utterances with such evidential marking have no bare alternatives to compete with, Gricean reasoning may give rise to different inferences than in a language like English. Exactly which inferences would be predicted depends on the particular evidential system and the semantic distinctions it encodes (Saratsli et al., 2020).

Acknowledgments

This work was partially supported by by the Israel Science Foundation (grant no. 1398/20 to Mira Ariel). We thank Mira Ariel and the ALPS Lab, especially Brandon Waldon, for helpful comments.

References

Aikhenvald, A. Y. (2014). The grammar of knowledge: A cross-linguistic view of evidentials and the expression of information source. In A. Y. Aikhenvald & R. Dixon (Eds.), *The grammar of knowledge: A cross-linguistic typology*. Oxford University Press.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*, 591–621.

Degen, J., Kao, J. T., Scontras, G., & Goodman, N. D. (2015). A cost and information-theoretic account of epistemic "must". *Poster Presented at CUNY 2015*.

Degen, J., Trotzke, A., Scontras, G., Wittenberg, E., & Goodman, N. D. (2019). Definitely, maybe: A new experimental paradigm for investigating the pragmatics of evidential devices across languages. *Journal of Pragmatics*, *140*, 33–48.

Faller, M. T. (2020). A possible worlds semantics for Cuzco Quechua evidentials. In C. Lee & J. Park (Eds.), *Evidentials and modals*. Brill.

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

Gisborne, N. (2010). *The event structure of perception verbs*. Oxford University Press.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

Grice, H. P. (1961). The causal theory of perception. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *35*, 121–152.

Korotkova, N. (2020). Evidential meaning and (not-)at-issueness. *Semantics and Pragmatics*, *4*, 1–26.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*, 335–342.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2012). Prosody and pragmatic inference: It looks like speech adaptation. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 647–653.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, *41*, 558–564.

Mandelkern, M. (2019). What 'must' adds. *Linguistics and Philosophy*, *42*, 225–266.

Matthewson, L. (2020). Evidence type, evidence location, evidence strength. In C. Lee & J. Park (Eds.), *Evidentials and modals*. Brill.

McCready, E. (2020). Testimony, trust, and evidentials. In C. Lee & J. Park (Eds.), *Evidentials and modals*. Brill.

McNally, L., & Stojanovic, I. (2017). Aesthetic adjectives. In J. O. Young (Ed.), *Semantics of aesthetic judgments*. Oxford University Press.

Muñoz, P. J. (2019). *On tongues: The grammar of experiential evaluation* (Doctoral dissertation). The University of Chicago.

Murray, S. E. (2020). A Hamblin semantics for evidentials and evidential questions. In C. Lee & J. Park (Eds.), *Evidentials and modals*. Brill.

Murray, S. E. (2021). Evidentiality, modality, and speech acts. *Annual Review of Linguistics*, *7*, 213–233.

Pogue, A., & Tanenhaus, M. K. (2018). Learning from uncertainty: Exploring and manipulating the role of uncertainty on expression production and interpretation. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 2266–2271.

Poortvliet, M. (2018). *Perception and predication: A synchronic and diachronic analysis of Dutch descriptive perception verbs as evidential copular verbs* (Doctoral dissertation). University of Oxford.

Roberts, C. (2004). Information structure in discourse. *Semantics and Pragmatics*, *5*, 1–69.

Saratsli, D., Bartell, S., & Papafragou, A. (2020). Cross-linguistic frequency and the learnability of semantics: Artificial language learning studies of evidentiality. *Cognition*, *197*, 104194.

Scontras, G., Tessler, M. H., & Franke, M. (2018). Probabilistic language understanding: An introduction to the rational speech act framework.

Simons, M. (2007). Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, *6*, 1034–1056.

Speas, M. (2018). Evidentiality and formal semantic theories. In A. Y. Aikhenvald (Ed.), *The Oxford handbook of evidentiality*. Oxford University Press.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, *25*, 359–393.

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