COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019) e12745

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12745

Evidential Strength of Intonational Cues and Rational Adaptation to (Un-)Reliable Intonation

Timo B. Roettger,[a] Michael Franke[b]

[a]*Department of Linguistics, Northwestern University & University of Cologne*
[b]*Institute of Cognitive Science, University of Osnabrück*

Received 8 December 2017; received in revised form 11 April 2019; accepted 15 April 2019

Abstract

Intonation plays an integral role in comprehending spoken language. Listeners can rapidly integrate intonational information to predictively map a given pitch accent onto the speaker's likely referential intentions. We use mouse tracking to investigate two questions: (a) how listeners draw predictive inferences based on information from intonation? and (b) how listeners adapt their online interpretation of intonational cues when these are reliable or unreliable? We formulate a novel Bayesian model of rational predictive cue integration and explore predictions derived under a concrete linking hypothesis relating a quantitative notion of evidential strength of a cue to the moment in time, relative to the unfolding speech signal, at which mouse trajectories turn towards the eventually selected option. In order to capture rational belief updates after concrete observations of a speaker's behavior, we formulate and explore an extension of this model that includes the listener's hierarchical beliefs about the speaker's likely production behavior. Our results are compatible with the assumption that listeners rapidly and rationally integrate all available intonational information, that they expect reliable intonational information initially, and that they adapt these initial expectations gradually during exposition to unreliable input. All materials, data, and scripts can be retrieved here: https://osf.io/dnbuk/

Keywords: Mouse tracking; Intonation; Prosody; Speech adaptation; Rational predictive processing; Probabilistic modeling

1. Introduction

One long-standing question of linguistic research is how listeners map a speech utterance onto intended meaning as rapidly and accurately as they do. This is not a trivial achievement because listeners have to integrate information from many different sources.

Correspondence should be sent to Timo B. Roettger, Department of Linguistics, Northwestern University, 2016 Sheridan Road, Evanston, IL 60208. E-mail: timo.b.roettger@gmail.com

The intended meaning of an utterance depends not only on *what* we say, that is, which words we use and how we combine them, but also on *how* we say them. For instance, we use intonation, that is, the modulation of fundamental frequency across the utterance (f_0), to encode sentence structure, illocutionary acts, and postlexical discourse relationships (e.g., Cruttenden, 1997; Cutler, Dahan, & Van Donselaar, 1997; Dahan, 2015; Gussenhoven, 2004; Ladd, 2008, among many others). Yet, despite its important role in human communication, we only have limited knowledge about how listeners' process intonation in order to recognize what a speaker intends to say.

A central concern for a theory of intonation-based intention recognition is how intonation is mapped onto discourse functions. Some authors have proposed a direct mapping of acoustic parameters onto discourse functions (e.g., Cooper, Eady, & Mueller, 1985; Fry, 1955), others have proposed a mediating level of abstract phonological representations. For example, in the Autosegmental Metrical model of intonation (e.g., Beckman & Pierrehumbert, 1986; Grice, 1995; Gussenhoven, 1984; Ladd, 2008; Pierrehumbert, 1980, among many others), an intonation contour is composed of tonal events located in structurally privileged positions. In their seminal work, Pierrehumbert and Hirschberg (1990) proposed that listeners directly identify and interpret these tonal events with regard to the relationship between the current utterance, the discourse context, and the assumed beliefs of the listener.

For example, information structure, that is, the way information is linguistically packaged to fit the context of the utterance and the knowledge state of the discourse participant, can be expressed by certain tonal events. In languages such as English and German, for instance, the position and form of a pitch accent, a tonal event co-occurring with a lexically stressed syllable, can signal a referent as discourse-given or discourse-contrastive (e.g., Baumann & Grice, 2006; Calhoun, 2007; Féry & Kügler, 2008; Ito & Speer, 2008; Pierrehumbert & Hirschberg, 1990; Rooth, 1992; Watson, Tanenhaus, & Gunlogson, 2008). In (1), a high rising pitch accent on the capitalized word can signal that this referent has to be interpreted in contrast to another, for example, an already established or contextually salient referent.

(1) a. Margarethe played the VIOLIN.
⇝ It was not the guitar.
b. MARGARETHE played the violin
⇝ It was not Sigfried.

Pierrehumbert and Hirschberg's model represented a point of departure for much research on possible intonational inventories and their relationship with intended meaning across a wide variety of languages (e.g., Jun, 2007, 2014, and references therein). An outstanding question, however, concerns the ubiquitous variability in how individual speakers map tonal events onto discourse meaning. A large body of evidence suggests that, even in the absence of contextual factors, speakers' intonational encoding of discourse functions varies; that is, speakers sometimes produce *categorically different* tonal events for the *same* discourse function and they sometimes produce one and the *same* tonal event for *different* discourse functions (e.g., for German: Cangemi, Krüger, & Grice

(2015), Grice, Ritter, Niemann, & Roettger (2017), for English: Chodroff & Cole (2018), Clopper & Smiljanic (2011), Ito, Speer, & Beckman (2004), Peppé, Maxim, & Wells (2000), Turnbull (2017), or for Berber: Roettger (2017)).

It becomes clear that assumed mappings between intonation and discourse meaning are regular but not necessarily deterministic (Bolinger, 1972; Hirschberg, 2002). Despite the large degree of variability, there are still (weak) statistical associations between intonation and discourse meaning in the sense that some intonation contours are more likely to be used to convey specific communicative functions than others. The question arises as to how listeners deal with this high level of variability when processing intonation.

It has been argued that the interpretation of intonation can still be seamlessly accomplished because the bottom-up perception of acoustic cues is heavily guided by probabilistic expectations about speaker production likelihoods, that is, how likely the speaker uses a particular intonational form in order to express a particular discourse function (Buxo-Lugo, 2017; Buxó-Lugo & Watson, 2016; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014b; Roettger, Mahrt, & Cole, 2019). For example, Kurumada et al. (2014b) showed that listeners' interpretation of a rising pitch accent depends on how reliably the speaker has used that pitch accent to express certain functions in the past.

Integrating intonation for intention recognition needs to involve a mechanism by which listeners evaluate the acoustic signal against expectations derived from other sources. This is evident from studies on real-time processing of intonational cues. Despite their inherent variability, listeners can rapidly integrate pitch accent information to anticipate a likely speaker-intended referent even before disambiguating lexical material is heard (e.g., Dahan, Tanenhaus, & Chambers, 2002; Ito & Speer, 2008; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014a; Roettger & Stoeber, 2017; Watson et al., 2008; Weber, Braun, & Crocker, 2006). These studies have demonstrated that listeners show anticipatory eye movements (or hand movements) when hearing an intonational event that allows them to predict the discourse status of an upcoming referent. For instance, Dahan et al. (2002) showed that listeners use pitch accent information to anticipate whether an upcoming referent has been already mentioned or is explicitly contrasted to a mentioned referent, with a high pitch accent being interpreted as evidence for a contrastive referent. However, listeners look to already mentioned referents when hearing a high pitch accent on a referent if that referent was less salient in the prior discourse. This suggests that a particular intonational cue (here the high pitch accent) is used to predict the speaker's intended meaning as a function of other discourse relationships. Watson et al. (2008) demonstrated that the predictive interpretation of a high pitch accent is compatible with both new and contrastive referents, again suggesting a flexible mapping of intonational form and discourse function.

In light of the variable nature of intonational form-function mappings, on one hand, and listeners' ability to rapidly integrate intonational information to anticipate referential intentions, on the other, it is important to examine the information integration process that maps the acoustic signal onto meaning. The goal of this paper is to shed more light on this process. In particular, we focus on two related issues: First, it is reasonable to assume that some acoustic cues are likely to be more informative (in the context in which they

occur) than others. We therefore propose a quantitative notion of evidential strength of intonational cues and test its predictions against novel empirical data. Second, we look more closely at the temporal development of listeners' online interpretation behavior over the course of an experiment. We present empirical data from two experiments in which listeners are exposed to various amounts of reliable or unreliable uses of intonational cues and compare the development of their interpretation behavior against the predictions of a model of conservative belief update from observation. In the following, we elaborate on both of these issues: evidential strength and adaptation.

1.1. Evidential strength

In the present paper, we spell out and critically test the idea that listeners predictively use intonational cues in a rational way, modeled here as Bayesian belief update. This approach, which is developed in detail in Section 2, presents us with a quantitative notion of evidential strength of intonational cues which is derived from general principles of rational information integration. The key idea is that listeners hold expectations about the speaker's differential likelihood of producing different utterances to express different meanings (cf. Franke & Jäger, 2016; Goodman & Frank, 2016). What matters for rational predictive interpretation are differences in the likelihood with which speakers are expected to produce a particular intonational contour when they wish to refer to one referent or another. By Bayes rule, a rational listener's posterior odds in favor of referent r_1 over r_2 after observing a (possibly partial) utterance u are calculated as the product of the likelihood ratio (how likely a speaker would produce u for r_i) and the prior odds (how likely a speaker would want to refer to r_i in the first place):

$$\underbrace{\frac{P(r_1 \mid u)}{P(r_2 \mid u)}}_{\text{posterior odds}} = \underbrace{\frac{P(u \mid r_1)}{P(u \mid r_2)}}_{\text{likelihood ratio}} \underbrace{\frac{P(r_1)}{P(r_2)}}_{\text{prior odds}}$$

Evidential strength of an intonational cue is equated with the likelihood ratio (Jaynes & Kempthorne, 1976; Jeffrey, 2002), that is, the amount of observational evidence in favor of an interpretation r_1 and against interpretation r_2 provided by a cue is given by how much more likely this cue would be produced for r_1 as compared to r_2.

A direct experimental measure of listeners' dynamically evolving posterior odds between two candidate interpretations can be obtained from mouse movements in a forced-choice decision task. Concretely, this paper adopts the linking hypothesis that posterior odds influence the final moment in time, relative to the unfolding speech signal, at which listeners' mouse movements turn towards the target to be chosen eventually. Numerous other experiments have demonstrated that the continuous uptake of sensory input and dynamic competition between simultaneously active representations is reflected in participants' hand or finger movements (Dotan, Meyniel, & Dehaene, 2018; Freeman & Ambady, 2010; Magnuson, 2005; Spivey, Grosjean, & Knoblich, 2005). This has also

been shown for intonational processing. Roettger and Stoeber (2017) have recently demonstrated that listeners integrate intonational information early on and move their mouse towards a likely target referent before they have processed disambiguating lexical information. These findings are in line with recent papers using mouse tracking to investigate the online processing of pragmatic inferences (e.g., Tomlinson, Bailey, & Bott, 2013; Tomlinson, Gotzner, & Bott, 2017).

1.2. Speech adaptation as adjusting evidential strength

Listeners' online interpretation is not static but can adapt flexibly to varying environments. Naturally, listeners' beliefs about the likely meaning a particular speaker may wish to convey when using a certain linguistic variant can change when they observe the speaker's actual production behavior. The observed behavior may diverge from the listeners' initial expectations. Language users have been repeatedly shown to adapt readily to their immediate local context in semantics/pragmatics (e.g., Grodner & Sedivy, 2011; Yildirim, Degen, Tanenhaus, & Jaeger, 2016), in syntax (e.g., Jaeger & Snider, 2013), in segmental speech categories (e.g., Bradlow & Bent, 2008; Kleinschmidt & Jaeger, 2015; Norris, McQueen, & Cutler, 2003), and intonation (Kurumada, Brown, & Tanenhaus, 2017; Kurumada et al., 2014b).

For example, Kurumada et al. (2017) exposed listeners to ambiguous intonation contours and contextually biased listeners' interpretation to either a contrastive or affirmative interpretation. They observed evidence for a shift in listeners' mapping of intonational cues onto respective interpretations. Similarly, Kurumada et al. (2014b) investigated listeners' online interpretation of intonational cues after pre-exposure to a speaker that either uses intonational cues in a natural and reliable way or in an unnatural and unreliable way. They showed that pre-exposure to unreliable input selectively blocked rapid intonational cue integration during the main experiment. Unfortunately, pre-exposure manipulation of cue validity gives only limited information about the temporal dynamics of listener adaptation when confronted with different frequencies of reliable or unreliable input. How do listeners adapt bit by bit to potential idiosyncrasies of a concrete speaker?

The studies presented in this paper try to answer this question by using a manipulation of the frequency of unreliable input within the experimental trials themselves and by looking at the temporal development of listener's predictive behavior over the course of the experiment. Building on the probabilistic notion of evidential strength of cues, we propose a hierarchical model of the listener's beliefs about the speaker's production behavior. Concretely, listeners start with initial beliefs about speaker production behavior which capture a priori expectations about natural language use. As listeners learn from observation how cues are produced to express one meaning or another, they update their expectations about the speaker's behavior. As a consequence of this dynamically evolving belief, the evidential value of any given cue might change. We, therefore, trace the model's predictions about the evidential strength of a cue, as it evolves under the input participants are faced with in our experiments. The model's dynamically evolving predictions of evidential cue strengths are then compared to our empirical data.

1.3. Overview

We present two experimental studies that address the questions of (i) how strongly various intonational cues impact listeners' interpretation early during utterance comprehension and (ii) how listeners dynamically adapt their online interpretation of intonational cues during exposure to either entirely reliable or occasionally unreliable form–function mappings. We use manual response dynamics as a window into listeners' posterior beliefs about the likely meaning of a partially observed utterance. Section 2 first introduces the probabilistic model for quantifying evidential strength and for capturing adaptation. Sections 3 and 4 introduce two experiments and discuss their results in light of the model's predictions. Section 5 discusses the results before Section 6 concludes.

2. Rational predictive processing of intonational cues

We propose a model of incremental and predictive interpretation of intonational cues. Following recent work in probabilistic pragmatics (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016), we assume that the listener's interpretation can be characterized as Bayesian inference. Under this assumption, the listener derives beliefs about likely interpretations of utterances—where we think of an utterance here as a possibly partial initial utterance fragment individuated by a particular intonational realization —from beliefs about the speaker's production behavior, that is, from beliefs about how likely the speaker is expected to use a particular utterance when wishing to express a given meaning.

Section 2.1 introduces a context model which captures the most relevant aspects of our experimental setup (see Sections 3 and 4). Section 2.2 uses this scenario to explain the notion of *evidential strength* associated with an utterance to quantify how much that utterance helps the listener decide between competing interpretation options. Quantified evidential strength will be related to the time course of disambiguation decisions, as measured by mouse tracking in the experiments discussed in Sections 3 and 4. Section 2.3 spells out general predictions of this models for these experiments. In order to capture flexible adaptation of interpretation under exposure to a particular pattern of speech, Section 2.4 looks at how the listener's beliefs about the speaker's production behavior are updated after every instance of observed speaker behavior.

2.1. Binary referential disambiguation

Margarethe and Sigfried are speakers of German. They are also scientists interested in the behavior of their new acquaintance, a fantastic creature called *the wuggy*. The wuggy likes to pick up things in its environment. Sigfried has been observing the wuggy on his own for a while. When Margarethe returns, she is curious to hear what she may have missed. Suppose that she asks a *topic question* like (2), which introduces a referent (the violin) as given for the subsequent discourse.

(2) Hat der Wuggy dann die Geige aufgesammelt?
Has the wuggy then the violin pick-up?
Did the wuggy then pick up the violin? ["violin?"]

Let us also assume that it is common knowledge between Margarethe and Sigfried that the wuggy picked up exactly one of two available objects: the violin mentioned in Margarethe's question "violin?" in (2) (the so-called *discourse-given referent*) and a pear which is not introduced into the discourse by Margarethe's question (the so-called *discourse-new referent*). Sigfried, who knows whether the wuggy picked up the violin or the pear, could answer Margarethe's question in many ways. Clearly, the sentences "violin" in (3) and "pear" in (4) are among his options.

(3) Der Wuggy hat dann die Geige aufgesammelt.
the wuggy has then the violin picked-up.
The wuggy then picked up the violin. ["violin"]

(4) Der Wuggy hat dann die Birne aufgesammelt.
The wuggy has then the pear picked-up.
The wuggy then picked up the pear then. ["pear"]

Even if we restrict attention to only these two sentence types, there are still several ways of realizing utterances of "violin" and "pear" in terms of their intonational manifestations. With a preceding topic question like "violin?" in (2), statements "violin" and "pear" in (3) and (4) are prototypically realized with different intonation contours in German (e.g., Féry & Kügler, 2008; Grice et al., 2017). After the polar topic question "violin?," the utterance "violin" affirmatively refers back to the discourse-given referent. This affirmation can be realized via a verum focus construction, which can intonationally manifest itself in the form of a rising-falling accent on the auxiliary verb *hat* (Engl.: *has*). We will henceforth refer to this contour as the VERB contour (see the orange line in Fig. 1A). As opposed to that, the answer "pear" corrects the topic question "violin?." It affirmatively mentions a discourse-new referent and is typically realized by an intonation contour with a prominent rising-falling pitch accent on the sentence object *Birne* (Engl.: *pear*). We will henceforth refer to this contour as the OBJECT contour (see blue line in Fig. 1A). Finally, although perhaps less typical, it is also possible to realize "violin" and "pear" with a *default intonation* which comes with a less prominent default accent on the object noun (see black line in Fig. 1A).

To say that there are prototypical intonational realizations of answers "violin" and "pear" after topic question "violin?" is not to say that listeners necessarily expect speakers to always exclusively produce the most prototypical pattern in any given situation. If intonational realizations are indeed variable as argued for in the introduction above, a listener who is aware of this holds a probabilistic belief $P_S(u \mid r, C)$ which encodes how likely it is, according to the listener, that the current speaker S realizes utterance u in question context C in order to express referential meaning r. The table in Fig. 1B gives an example of such a belief about probabilistic speaker behavior. The example assumes, as we will throughout this paper, that the listener expects a

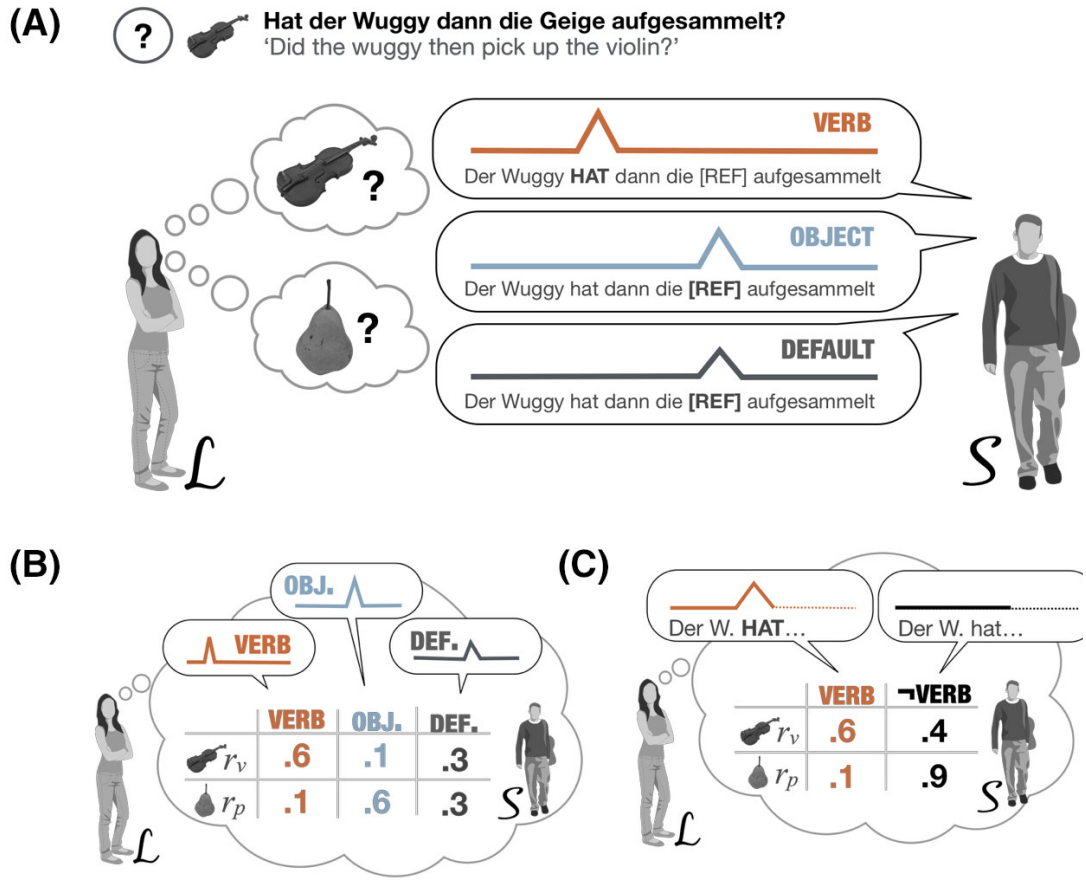

Fig. 1. (A) Context of interpretation. After a topic question introducing the violin into the discourse, the speaker realizes an utterance of the form "Der Wuggy hat dann die [REF] aufgesammelt" (Engl.: *The wuggy then picked up the [REF]*) using different intonation patterns. The referent is either one of two salient contextually given referents, for example, a violin or a pear. (B) Example of a listener's belief about the speaker's conditional production probabilities. For each referent, a probability is assigned to the intonational realization of a lexically appropriate utterance. (C) The listener's beliefs after observing a partial utterance derived from the beliefs shown in (B).

semantically true utterance (i.e., when the speaker wants to express that the wuggy picked up the violin he will not mistakenly utter the sentence "pear" in (4)) and that the listener expects the realization of different intonational contours to be more or less likely. For exposition purposes, this example encodes a listener's expectation that a VERB contour is realized with probability .6 when uttering sentence "violin" and with probability .1 when uttering sentence "pear" in the context of question "violin?". In the next section, we will come back to the issue of which listener beliefs about speaker production are justifiable or reasonable.

2.2. Evidential strength of intonational cues

The beliefs of Margarethe in Fig. 1B assign probabilities to intonational realizations of *complete* utterances of Sigfried. From these, we can derive Margarethe's beliefs about partial utterances. Margarethe's belief that Sigfried will realize partial utterance u_{partial} in context C to express meaning r, is given by the sum of all complete utterances u that start with u_{partial}:

$$P_S(u_{\text{partial}} \mid r, C) = \sum_{u-\text{a complete utterance starting with } u_{\text{partial}}} P_S(u \mid r, C)$$

Fig. 1C shows Margarethe's beliefs, derived from those in Fig. 1B, after a partial utterance "Der Wuggy hat..." ("The wuggy has...") with either a prominent pitch accent on the auxiliary or not. The latter intonational realization is compatible with two complete utterances listed in Fig. 1B, namely the OBJECT and the DEFAULT contour, so that the conditional probabilities for what we denote as—VERB in Fig. 1C are computed as the sum of the two corresponding probabilities for complete utterances.

Regardless of whether u is a partial or a complete utterance, we can quantify the evidential strength of u as the extent to which an observation of u changes the listener's beliefs about the relative probability (so-called *odds*) of the competing interpretations. Following recent work in probabilistic pragmatics (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016), we assume that the listener's interpretation of an observed utterance u follows Bayes rule to assign probabilities to possible interpretations, like so:

$$\underbrace{P_L(r \mid u, C)}_{\text{posterior}} \propto \underbrace{P_S(u \mid r, C)}_{\text{likelihood}} \underbrace{P(r \mid C)}_{\text{prior}},$$

where $P(r \mid C)$ is the listener's prior degree of belief that the speaker wants to express (referential) meaning r. Since there are only two possible referents at stake, we can look at Margarethe's *posterior odds* in favor of referent r_1 over r_2 after observing a (possibly partial) utterance u. These are calculated, by Bayes rule, as the product of the likelihood ratio (how likely a speaker produces u for r_i) and the prior odds (how likely a speaker refers to r_i):

$$\underbrace{\frac{P_L(r_1 \mid u, C)}{P_L(r_2 \mid u, C)}}_{\text{posterior odds}} = \underbrace{\frac{P_S(u \mid r_1, C)}{P_S(u \mid r_2, C)}}_{\text{likelihood ratio}} \underbrace{\frac{P(r_1 \mid C)}{P(r_2 \mid C)}}_{\text{prior odds}}$$

All else equal, if utterance u with its specific intonation contour is more likely to be produced for r_1 than for r_2, an observation of u would shift the listener's beliefs towards r_1 and away from r_2. Observing u would therefore be *observational evidence* in favor of r_1

relative to r_2 (Jaynes & Kempthorne, 1976; Jeffrey, 2002). We refer to the likelihood ratio $\frac{P_S(u|r_1,C)}{P_S(u|r_2,C)}$ as the *evidential strength* of (partial) utterance u in favor of r_1 over r_2 and say that u provides *positive evidence* in favor of r_1 over r_2 when $\frac{P_S(u|r_1,C)}{P_S(u|r_2,C)} > 1$. Notice that, in order to determine how much evidence an observation of u carries in favor of hypothesis r_1, the prior is irrelevant; it only matters whether observation u would be more likely produced for r_1 than for r_2.

Partial utterances can provide positive evidence for one referential interpretation over another, simply in virtue of their intonation contours. In the example from Fig. 1C, for instance, an observation of VERB has an evidential strength of $\frac{.6}{.1} = 6$ in favor of the meaning r_v, indicating that the wuggy picked up the violin. This means that, no matter what Margarethe believed a priori about the probability of referential meanings, after observing this partial utterance, she would rationally adjust her prior odds by a factor of 6 in favor of interpretation r_v. That is a rather noteworthy shift, going from prior odds to posterior odds. In contrast, an observation of partial utterance type ¬ VERB in Fig. 1C) has an evidential strength of $\frac{.9}{.4} = 2.25$ in favor of the meaning that the wuggy picked up the pear. Two things are interesting here. For one, the two partial utterances in this example provide positive evidence for different interpretation hypotheses. For another, the evidential strength associated with these cues differs quantitatively as well. In later Sections 3 and 4, we will link evidential strength quantitatively to the time course of disambiguation. The stronger an early intonational cue is, the sooner listeners will choose the interpretation for which the cue provides positive evidence. (Indeed, we will assume that prior odds are sufficiently close to 1 to be negligible.) What the simple example presented here demonstrates is that the evidential strength of an intonational realization, as defined here, does not hinge on whether a salient cue (here a pitch accent) is present or absent but on the differential likelihood of realizing the currently observed intonational pattern with a higher probability for one referential meaning rather than another.

2.3. Generalizing the example

The example given so far only used by and large arbitrary fixed numbers to characterize the listener's beliefs about speaker production likelihoods. As modelers, we should not commit to a single set of such numbers, but derive more general predictions based on a wide range of plausible parameter values. Therefore, let the listener's beliefs about relevant speaker production likelihoods be given as in the following table:

	Partial utterance	
Referent	V	$\bar{V}$
r_v	p_V	$1 - p_V$
r_p	ε_V	$1 - \varepsilon_V$

As before, the focus is on a situation in which the context question is "violin?" in (2) and the listener has already heard the beginning of the utterance "Der Wuggy. . ." ("The wuggy. . ."). There are only two prominent intonational realizations of the next lexical item, which the listener knows to be *hat* (Engl.: *has*): It either carries a pitch accent

(option V) or it does not (option $\overline{V}$). Here, p_V is the probability of producing V when wishing to refer to the given referent r_v (the violin), and ε_V is the probability of producing V when wishing to refer to the discourse-new referent r_p (the pear). In the following, we will impose a set of what we believe are plausible constraints on these two parameters—plausible in the sense that they agree with our (modelers') intuitions about the frequencies of intonational variations among German speakers. From these general constraints, we will then derive predictions about the ordering of evidential strength associated with a pitch accent on the auxiliary and the absence thereof.

Assumptions about typical intonational realizations of whole sentences "violin" (3) and "pear" (4) after context question "violin?" (2) by German speakers translate into expectations about likelihoods of partial utterances of an initial sentence fragment "Der Wuggy hat ..." ("The wuggy has ..."). Realizing a pitch accent on the auxiliary (option V) in a sentence like "pear" (4), which refers to the discourse-new referent r_p, is a genuine violation of typical uses of this intonation contour in German. Production studies on German confirm this intuition (Turco, Dimroth, & Braun, 2013). We, therefore, assume that ε_V is rather small but, as errors are conceivable in principle, still positive: $0 < \varepsilon_V$. In contrast, the realization of a pitch accent on the auxiliary in sentence "violin" (3) (option V) to refer to the discourse-given referent r_v is clearly not as unnatural as it is for "pear", so that $\varepsilon_V < p_V$ (Grice, Lohnstein, Röhr, Baumann, & Dewald, 2012; Turco et al., 2013). Finally, a default intonation of "violin" (3) is also a natural possibility. It is not important here to commit to whether p_V is bigger or smaller than $1 - p_V$. What is important for us, and also rather uncontroversial, is that since default intonation of "violin" is not unnatural, but a verb contour in "pear" is, we expect that $1 - p_V < 1 - \varepsilon_V$. Consequently, we end up assuming the following order relations between likelihood parameters: $0 < \varepsilon_V < p_V < 1 - \varepsilon_V$.

With these assumptions in place, the main observations obtained for the example from Fig. 1 hold more generally. First, as $\varepsilon_V < p_V$, we have $\frac{p_V}{\epsilon_V} > 1$, so that the presence of an early pitch accent on the auxiliary, represented here as V, provides positive evidence for the predictive interpretation that the speaker likely wants to pick out the discourse-given referent r_v. Second, as $1 - \varepsilon_V > 1 - p_V$,we have $\frac{1-\epsilon_V}{1-p_V} > 1$, so that we predict that the absence of a pitch accent, represented here as $\overline{V}$, provides early positive evidence for r_p. Finally, the quantitative picture endorsed here allows for ordinal comparison: we expect that V provides stronger evidence for r_v than $\overline{V}$ provides for r_p.

Proposition 1. On the assumption that $0 < \varepsilon_V < p_V < 1 - \varepsilon_V$, the evidence in favor of the given referent provided by a pitch accent on the verb is higher than the evidence in favor of the new referent provided by the absence of such a pitch accent: $\frac{P(V|r_v)}{P(V|r_p)} > \frac{P(\overline{V}|r_p)}{P(\overline{V}|r_v)}$.

Proof. We assume that $0 < \varepsilon_V < p_V < 1 - \varepsilon_V$ and need to show that:

$$\frac{p_V}{\epsilon_V} > \frac{1 - \epsilon_V}{1 - p_V}.$$

With $x = \varepsilon_V$, $x' = 1 - p_V$ and $c = p_V - \varepsilon_V = (1 - \varepsilon_V) - (1 - p_V)$, we can rewrite this like so:

$$\frac{x + c}{x} > \frac{x' + c}{x'}.$$

The inequality follows from the observation that $f(x) = \frac{x+c}{x}$ is strictly monotone decreasing and concave when $x, c > 0$. These conditions are met by our assumption. For monotonicity, note that $f'(x) = -\frac{c}{x^2} < 0$ for $x, c > 0$. For concavity, note that $f''(x) = \frac{2c}{x^3} > 0$ for $x, c > 0$.

Proposition 1 allows to make ordinal predictions about the time course of interpretation in the experiments reported in Sections 3 and 4. In a nutshell, the presence of an early pitch accent in a VERB contour provides a stronger cue than its absence which, however, also carries useful information and may thus facilitate predictive interpretation.

2.4. Rational adaptation to observed input

So far, we have described a listener with a rather specific, probabilistic belief about the speaker's production behavior. In the following, we are interested in extending this model to capture how Margarethe would change her beliefs about Sigfried's behavior, when she observes repeatedly how Sigfried actually chooses utterances to convey certain meanings. Unfortunately, it is not possible to say in general how Margarethe should update her beliefs after an observation of Sigfried's behavior when all we have is a representation of her beliefs as in Fig. 1. The following illustration makes this argument more accessible: Suppose Margarethe's beliefs about Sigfried's utterances V and $\overline{V}$ are as in Fig. 1. She believes that, with probability .6, Sigfried realizes an utterance of "violin" with a VERB contour. Abstractly, Sigfried's choice of expression is like the flip of a coin with a bias towards one option over the other. For exposition purposes, let us say the coin is biased toward heads by .6. If Margarethe now observes Sigfried produce "violin" with a VERB contour in the relevant context, this is like observing an outcome of heads after a single coin flip. If Margarethe believed that this outcome had a probability of .6 initially, what should she believe after her observation?—We cannot really say. It is clear that Margarethe's new beliefs should shift in a particular direction: The probability of the VERB contour should go up. But for how much? Intuitively speaking, the problem is that the numerical information we have at hand is insufficient because we do not know how confident Margarethe is that Sigfried's utterance probability is exactly .6. If she is very confident, a single observation would change these probabilities only very little, say to .601. If she is not confident at all, a single observation might change her beliefs much more, say to .7.

To overcome this problem, a standard solution is to consider a hierarchical, so-called Dirichlet-Multinomial model. The key idea is that rather than just entertaining one representation of the speaker's probabilistic production behavior, we model Margarethe as entertaining a *higher-order belief* about all possible ways in which Sigfried could probabilistically choose utterances to encode meanings. This approach is visualized in Fig. 2.

Margarethe's higher-order beliefs are captured conveniently in a matrix of so-called Dirichlet weights (see Fig. 2A). The matrix of Dirichlet weights defines a full probability distribution over all ways in which Sigfried could probabilistically realize utterances for different meanings. For instance, for the Dirichlet weights given in Fig. 2A, Margarethe's expectation of the probability of Sigfried producing V for r_v is $\frac{60}{100} = 0.6$, but she also considers it possible that it is .7.

The main benefit of this approach is that it captures belief updates from observation very elegantly. Essentially, a single observation of Sigfried's behavior changes Margarethe's higher-order beliefs by a simple increment of the cell in the matrix of Dirichlet weights that corresponds to her observation (see Fig. 2B). For example, suppose Margarethe observes an utterance of "violin" with VERB contour which expresses meaning r_v. We simply add one to the entry in the relevant cell in the matrix of Dirichlet weights (see Fig. 2B), resulting in an updated expectation of the probability of producing V for r_v as $\frac{61}{101} \approx 0.604$, an increment by ca. 0.04. If, instead Margarethe's beliefs had been modeled by Dirichlet weights $\langle 6, 4 \rangle$, the increment would have been larger: $\frac{7}{11} \approx 0.64$. Indeed, the absolute amount of non-normalized weights (the sum of numbers for a given

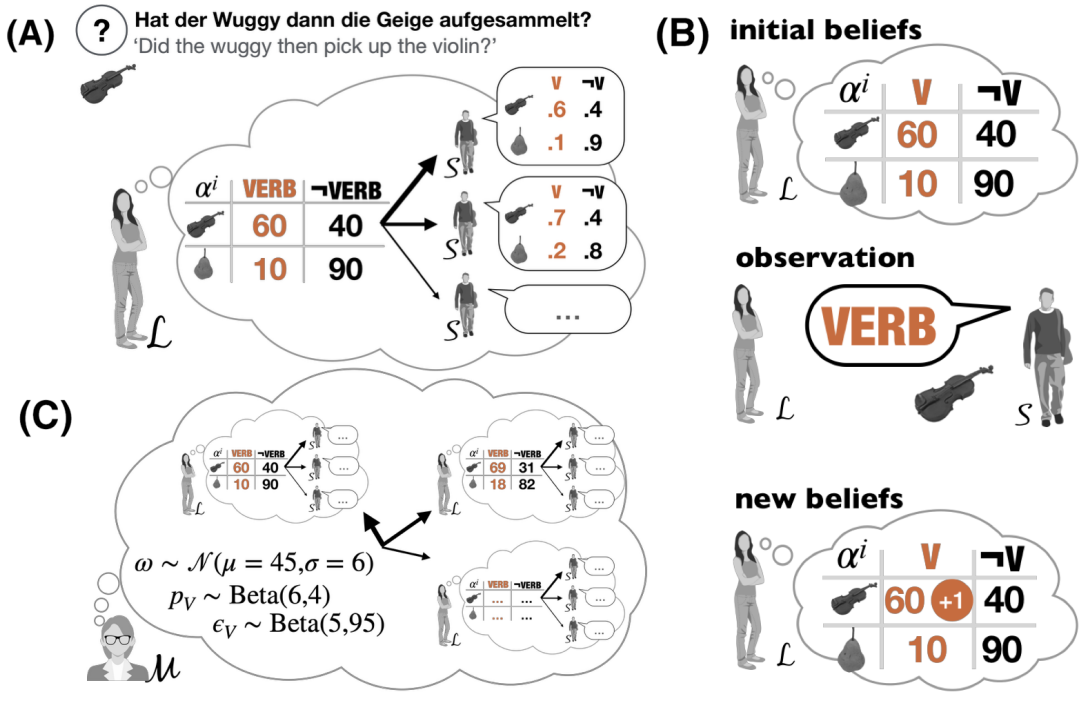

Fig. 2. (A) A listener with a Dirichlet-multinomial belief that defines a probability distribution over beliefs about probabilistic speaker behavior. (B) Illustration of learning from observation in the Dirichlet-Multinomial model. (C) Illustration of the modeler's beliefs about the listener's Dirichlet weights.

row) determines the impact a single observation has on the difference between prior and posterior expectation.

Appendix A spells out the model and its properties in more formal detail. It also shows how the results obtained in Section 2.3 and Proposition 1 for Margarethe's nonhierarchical beliefs, are all conserved under this hierarchical extension. Readers who are less familiar with the concepts and notational conventions used in Appendix A will not miss much of substance when they read on with an intuitive understanding of the model, roughly as conveyed in Fig. 2.

We will use this general model of belief dynamics in Sections 3 and 4 to derive specific predictions for the temporal development of predictive processing of intonational information. Given the design of our experiments, we focus on the contrast between V and $\overline{V}$ as before, that is, the difference between a partial utterance with and without an early pitch accent on the verb. To encode assumptions about listener beliefs about speaker likelihoods which resonate with intuitions about normal or usual language use in German, we will consider yet another layer of uncertainty on top of the Dirichlet-Multinomial belief model considered so far (see Fig. 2C). In this doubly hierarchical model, the listener has higher-order uncertainty about production likelihoods captured by a single matrix of Dirichlet weights, and we, as modelers, have uncertainty about these higher-order beliefs, that is, uncertainty about which matrix of Dirichlet weights best characterizes listeners' higher-order beliefs (see Fig. 2C). To efficiently capture modeler's uncertainty, we will assume the relevant Dirichlet weights to be:

	VERB	¬VERB
r_v	$p_V\omega$	$(1 - p_V)\omega$
r_p	$\varepsilon_V\omega$	$(1 - \varepsilon_V)\omega$

Here, p_V and ε_V are as before and $\omega > 1$ is a factor that determines, intuitively speaking, the plasticity of beliefs, that is, the impact a single observation has on the listener's beliefs about speaker likelihoods. The higher ω, the more confident the listener is assumed to be and so beliefs will change less rapidly when observing a speaker's actual behavior. We here assume the following priors over these model parameters:

$$\omega \sim \mathcal{N}(\mu = 45, \sigma = 6) \quad p_V \sim \text{Beta}(6, 4) \quad \epsilon_V \sim \text{Beta}(5, 95)$$

These priors are in line with our previous categorical assumption that $0 < \varepsilon_V < p_V < 1 - p_V$ in that the prior probability sampling of a pair $\langle p_v, \varepsilon_V \rangle$ which does not conform to this constraint is around $8e^{-6}$ (estimated by Monte Carlo sampling).

This model lets us derive concrete predictions regarding how listener's predictive interpretation of early intonational cues changes over the course of an experiment. The exact predictions of this model for the experiments reported here will be discussed in Sections 3.2 and 4.2 when all the necessary details of the experiments have been introduced.

3. Experiment 1

The following experiment was preregistered on July 5, 2017, prior to data collection. The preregistration file can be retrieved alongside all materials, raw data, and corresponding analysis scripts from https://osf.io/dnbuk/.

3.1. Method

3.1.1. Participants and procedure

Sixty native German speakers participated in the study. All participants had self-reported normal or corrected-to-normal vision and normal hearing (30 male, 30 female, M_{age} = 25.3 (SD = 3.1)).

Participants were seated in front of a Mac mini 2.5 GHz Intel Core i5. They controlled the experiment via a Logitech B100 corded USB Mouse. Cursor acceleration was linearized and cursor speed was slowed down (to 1,400 sensitivity) using the CursorSense© application (version 1.32). Slowing down the cursor ensured that motor behavior was recorded as the acoustic signal unfolded, resulting in a smooth trajectory from start to target (Kieslich, Schoemann, Grage, Hepp, & Scherbaum, 2019).

Participants were told about a fantasy creature called "wuggy," which picks things up. There were 12 different objects that the wuggy could pick up (bee, chicken, diaper, fork, marble, pants, pear, rose, saw, scale, vase, and violin; see Fig. 3C). Each trial exposed participants first to a context screen, which was shown for 2,500 ms and provided a specific discourse context (see Fig. 3A). The question screen displayed an uninformative image of a headphone. Concretely, participants heard either a *topic question* like (2), repeated here from above, which introduces a referent as given into the discourse, or they heard the *neutral question* in (5).

(2) Hat der Wuggy dann die Geige aufgesammelt? [topic question]
Did the wuggy then pick up the violin?

(5) Was ist passiert? [neutral question]
What happened?

Following the question screen, participants saw a response screen with two visually presented response alternatives, each depicting one referent in the upper left and right corner, respectively (left/right placement of target vs. competitor response alternatives was counterbalanced within participants and items). After 1,000 ms, a yellow circle appeared at the bottom center of the screen. When participants clicked on the yellow circle, they initiated playback of an audio recording of a statement specifying which object was picked up, for example, (3) or (4), repeated from above.

(3) Der Wuggy hat dann die Geige aufgesammelt.
the wuggy has then the violin picked-up.
The wuggy then picked up the violin.

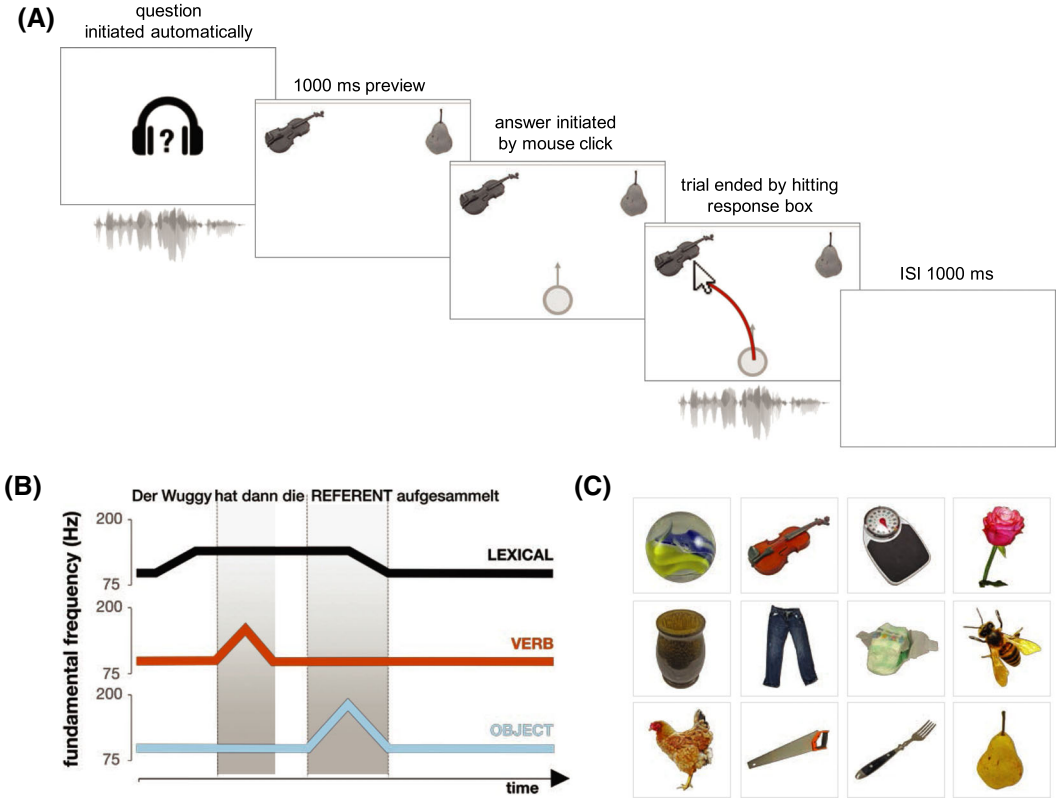

Fig. 3. (A) Schematic depiction of experimental trials. On the question screen, participants heard the context-setting question. After a 1,000 ms preview of target and competitor referents, the initiation button was displayed at the bottom center of the screen. Upon clicking the initiation button, listeners started the audio playback of the response sentence, indicating the target referent. The trial ended when hitting the response box surrounding one of the referents. Interstimulus intervals were 1,000 ms. (B) Schematic f_0 contours and average temporal landmarks for the resynthesis of the three critical intonation contours . (C) The 12 visual referents used in the experiment.

(4) Der Wuggy hat dann die Birne aufgesammelt.
the wuggy has then the pear picked-up.
The wuggy then picked up the pear then.

Statements were acoustically manipulated to exhibit three different intonation contours (see Fig. 3B), namely the VERB contour and the OBJECT contour, as introduced in Section 2.2, as well as a hat pattern, characterized by a rise in f_0 on the subject, a high plateau, and a subsequent fall in f_0 towards the sentence object. This contour is a rather neutral contour that can be used to express out-of-the-blue statements and is here used for our baseline LEXICAL disambiguation. Because the question (5) does not introduce a referent into the discourse, listeners have to wait for the acoustic information of the noun

itself in the LEXICAL condition. All possible statements (n = 12) came with these three intonation contours, resulting in 36 different target sentences.

Participants were instructed to move their mouse immediately upwards after clicking the initiation button and to choose the respective response alternative as quickly as possible. If they did not initiate their movement immediately (i.e., within 350 ms), they automatically received feedback that reminded them to do so. This time pressure ensured that participants began their mouse movement before the onset of relevant acoustic information, which enables distinguishing properties in the acoustic signal to influence the continuous motor output during its movement (Fischer & Hartmann, 2014; Hehman, Stolier, & Freeman, 2015). After each response selection, the screen was left blank for a 1,000 ms interstimulus interval. Prior to the experimental trials, participants familiarized themselves with the paradigm during 16 practice trials.

There were two experimental groups. The reliable speaker (RS) group was only exposed to intonation patterns that matched the discourse context and the lexical information in each sentence, as described above. Listeners could therefore rely on the systematic mapping of intonational form (pitch accent position) and function (the respective discourse status of the referent). As opposed to that, the unreliable speaker (US) group was sometimes exposed to mismatching intonation; that is, one out of three VERB/OBJECT trials was a mismatch. A mismatch occurs when, in the context of a topic question like (2), the speaker surprisingly uses a pitch accent on the OBJECT to indicate a discourse-given referent; or surprisingly uses a pitch accent on the VERB to indicate a discourse-new referent. Occasional mismatch leads to a scenario in which listeners cannot fully rely on the speaker's form–function mappings. Concretely, participants were exposed to 12 blocks of eight stimuli each. In the RS group, each block contained two OBJECT trials, two VERB trials, and four LEXICAL trials, resulting in 96 trials overall. Each block in the US group was the same except that there were only two LEXICAL trials and, additionally, two unreliable mappings (one with a mismatching OBJECT contour and one with a mismatching VERB contour). In sum, the US group received additional unreliable trials but less control trials than the RS group. Unreliable and reliable trials were randomly interspersed in the US group.

Each participant was randomly assigned to one reliability group. There were 30 participants in each group. Item pairs and their combination with intonation contour were pseudo-randomized for each block. Order of trials within each block and order of blocks were randomized for each participant.

3.1.2. Material

Visual stimuli were taken from the BOSS corpus (Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010). There were two sets of acoustic stimuli: questions providing a discourse context presented on the question screen and statements triggering participants' responses on the response screen. Thus, there was one question and one statement corresponding to each object.

Acoustic stimuli were recorded by a trained phonetician in a sound-attenuated booth with a headset microphone (AKG C420) using 48 kHz/16-bit sampling. To ensure that the three different conditions exhibit the same temporal characteristics for each sentence

(i.e., the lexical information of the referent becomes available at the same time), sentences were manipulated and resynthesized using Praat (Boersma & Weenink, 2016) applying the following procedure.

We took the original stimuli produced with a VERB contour as a point of departure because they can easily be resynthesized into the other two intonational patterns without creating mismatching acoustic information. We took one prototypical statement produced with a VERB contour and isolated the first part of the sentence ("Der Wuggy hat", Engl.: *the wuggy has*). We refer to this single part as the "left splice." For each individual sentence, we isolated the rest of the sentence after "hat" (e.g., "dann die Birne aufgesammelt", Engl.: *collected the pear then*). We refer to these parts as the "right splices." The midpoint of the voiceless stop closure of "hat" was chosen as the point to splice the two parts of the signal together. The single left splice was now concatenated with each right splice, respectively, resulting in 12 different base sentences, exhibiting the same temporal landmarks up to "hat."

In the next step, we manipulated the duration of "hat" and the stressed syllable of each referent (e.g., "BIRne", Engl.: *pear*). In order to ensure that the baseline enables the perception of an accent either on "hat" or on the referent, we reduced the duration of "hat" by a factor of 0.7 and increased the duration of the stressed syllable of the referent by a factor of 1.2. The resulting manipulations were taken as instances of the VERB condition and were further processed for the resynthesis of OBJECT and LEXICAL stimuli.

For the LEXICAL contour, we decreased the intensity of "hat" and increased the intensity of the stressed syllable of the subject ("WUggy") as well as the referent (e.g., "BIRne") in order to facilitate the impression of accents on these constituents. We then changed the f_0 contour as follows: We included a rise in f_0 (30 Hz) starting at the word onset of the subject ("Wuggy") and ending at the end of its stressed syllable. Following the rise, f_0 remained high until the end of the stressed syllable of the referent and fell towards the end of the word (30 Hz). The rest of the utterance remained low, resulting in a hat pattern, commonly observed for neutral statements in German (Grice et al., 2017).

For the OBJECT condition, we decreased the intensity of "hat" and increased the intensity of the stressed syllable of the referent (e.g., "BIRne" "pear"). We then changed the f_0 contour as follows: We flattened the rise in pitch on "hat" and included a high rise in f_0 (50 Hz) starting at the word onset of the referent and reaching its maximum at the end of the stressed syllable. Following the rise, f_0 fell (50 Hz) towards the end of the stressed syllable of the referent. f_0 for the rest of the utterance remained low, resulting in a rise–fall on the accented referent, commonly observed for contrastive focus in German (Grice et al., 2017).

3.1.3. Data analysis

The screen coordinates of the computer mouse were sampled at 100 Hz using the mousetrap plugin (Kieslich & Henninger, 2017) implemented in the open source experimental software OpenSesame (Mathôt, Schreij, & Theeuwes, 2012). Trajectories were processed with the package `mousetrap` (Kieslich & Henninger, 2017) using `R` (R Core Team, 2017).

There was a total of 96 target trials for the RS group. For the US group, we only analyzed the 72 target trials with reliable mappings between discourse context and

intonation. By design, three factors are relevant for the analysis. The factor CONDITION encodes whether the stimulus sentence was realized with a pitch accent on the VERB, on the OBJECT, or whether the discourse was neutral, leaving only the LEXICAL material disambiguating the two interpretations. Conditions are triggered by either the neutral question in the LEXICAL condition or by the topic question in the VERB and OBJECT conditions. The factor GROUP is a between-subjects contrast, encoding whether input was always RELIABLE or occasionally UNRELIABLE. Finally, to test the temporal development of anticipatory behavior over the course of the experiment, we included the scaled numerical predictor BLOCK.

In order to link manual response dynamics to listeners' dynamically unfolding posterior beliefs about likely interpretations (see the next Section 4.2), we look at the moment in time relative to the unfolding speech signal at which a mouse trajectory starts to migrate uninterruptedly towards the eventual interpretation choice. We define the *turn towards the target* (TTT) as the latest point in time at which the trajectory did not head towards the target.[1,2]

3.2. Model predictions

As detailed in Section 2, we expect that a rational incremental interpreter who predictively uses even weak probabilistic information from early intonational cues and who holds natural (higher-order) beliefs about speaker production likelihoods will conceive stronger evidence in the VERB condition than in the OBJECT condition. We predict no early disambiguating evidence before lexical disambiguation in the LEXICAL condition. These processing differences are expected to show early during the course of the experiment in both RS and US groups.

To link these model predictions to a concrete empirical measure, we assume that the TTT measure reflects the listener's uncertainty about which referent is meant by the speaker. The TTT measure will be lower—the decision will be faster and more confident—if the posterior odds of the target are higher earlier during sentence processing: the more certain participants are at the current stage, the more likely it is that their mouse trajectory gravitates towards the target already. This assumption is in line with the general idea of ballistic accumulator models (e.g., Ratcliff & McKoon, 2008), according to which evidence in favor of a choice or hypothesis accumulates stochastically over time and results in execution if a critical mass is met. According to this linking hypothesis, the TTT measure is a strictly decreasing function of the posterior odds in favor of the target referent. We cannot offer a definite theory or commit to a precise mapping from posterior odds to TTT, but it is clear that the latter must have a finite lower bound to which it converges from above as posterior odds grow to infinity. This is because even if a cue is fully disambiguating, the TTT cannot happen before the cue is perceived and processed. Thus, we suggest an exponential decay function as an approximate link function:

$$\text{TTT} \sim \exp(1 - \text{posterior odds}).$$

This approximate link function does not allow fine-grained quantitative predictions, but it does allow for nontrivial qualitative predictions concerning the development of interpretation behavior over the course of the experiment. Each block of trials in the RS group will increment the Dirichlet weights, which characterize the (higher-order) beliefs of the listener (see Section 2.4), with the summands given in Table 1a, and those in the US group with the summands in Table 1b.

Given this knowledge about what participants see during the experiment, as well as our modeler's priors over Dirichlet weights of our listener model, we can compute expectations regarding the temporal development of TTT values. Fig. 4 shows a summary of these predictions, in terms of how likely the assumptions spelled out so far make it that the measured TTT values will increase, stay constant or increase over the course of the experiment in different experimental conditions. Appendix A explains in detail how these categorical predictions were derived.

In other words, we expect that for the VERB condition in the RS group there will almost surely not be a noticeable difference in the TTT measure between the first and last block of the experiment. With a very low but still positive probability, we expect a facilitatory effect, that is, a decrease in TTT measurements as the experiment proceeds. This is because we expect the evidential strength of the pitch accent on the auxiliary to be rather strong already at the outset of the experiment, so that further strengthening via observational learning will likely have little additional effect. In the OBJECT condition of the RS group, we mostly expect a facilitatory effect of experience, that is, a decrease in TTT measurements. We also put a non-negligible probability mass on the possibility that TTT measurements remain constant. Intuitively, the absence of a pitch accent on the verb is initially believed to be a weak cue. These beliefs are then incrementally updated upon reinforcing evidence that this cue is informative for the interpretation of the (partial) utterance.

When we turn to the US group, we predict that a slowdown, that is, an increase in TTT measure, is most likely for the VERB condition, since observing unreliable uses of a

Table 1
Increments of Dirichlet weights of the listener's higher-order beliefs after one block of experimental trials in Experiment 1

(a) Reliable speaker group

	V	$\overline{V}$
r_v	2	0
r_p	0	2

(b) Unreliable speaker group

	V	$\overline{V}$
r_v	2	1
r_p	1	2

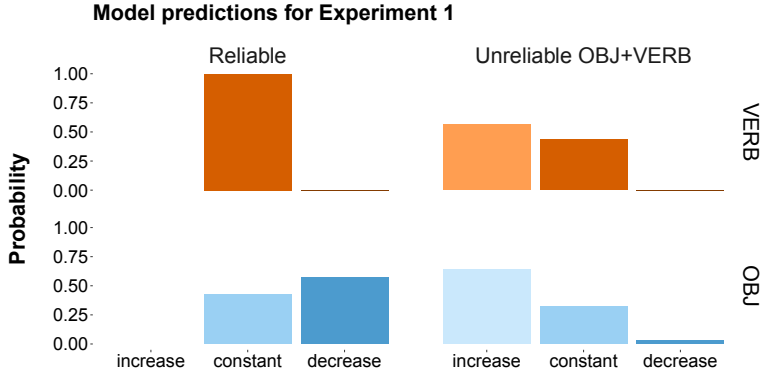

Fig. 4. Categorical predictions about the temporal development of the TTT measure throughout Experiment 1. The plots show the probability with which we should expect that the TTT measurements decrease, stay constant, or increase in different experimental conditions over the course of the experiment (given the assumptions spelled out in the main text).

pitch accent on the auxiliary will quickly reduce the initially high evidential strength associated with it. Our predictions for the OBJECT condition in the US group are less specific. Any development is compatible with the assumed prior beliefs over model parameters, but we predict that it is least likely to see a decrease in TTT measurements, while we expect constancy to be most likely.

3.3. Results

Following preregistered protocol, the whole data of participants were excluded whenever they (a) exhibited more than 10% errors, (b) exhibited movement behavior violating instructions in more than 15% of the trials, or (c) exhibited initiation times above 350 ms in more than 15% of the trials. For each of the exclusion criteria, we had to exclude one participant.

Trials with initiation times greater than 350 ms (1.5%) and incorrect responses (0.3%) were discarded on a trial-by-trial basis. Additionally, trials that exhibited movement behavior violating instructions were discarded, too (1.1%).

3.3.1. Descriptive assessment of trajectories

Fig. 5 displays the time and space normalized trajectories for three subgroups of trials. Listeners' mouse trajectories during expected form–function mappings (VERB-given and OBJECT-contrastive) are characterized as follows. Listeners move the cursor up in a straight line gravitating towards the midpoint of the screen and upon accumulating enough evidence from the acoustic signal, listeners turn towards the target. In response to the unexpected patterns presented in the US group (right most panel), listeners initially move up straight and then move towards the competitor before eventually curving towards the target. This detour is particularly pronounced in the verb condition with a strong attraction

towards the competitor. These patterns suggest that listeners interpret the available pitch accent information in the signal as informative for reference resolution and then correct their decision process on the fly upon hearing the lexically disambiguating information later in the signal.

Because of the form of the trajectories in the reliable trials (initially gravitating towards the middle and then smoothly turning towards the target) and due to the stimuli spanning large temporal windows, it is informative to investigate properties of the trajectories relative to temporal landmarks in the acoustic stimuli. We are interested in the point in time when listeners' manual movements indicate that the available evidence in the signal makes the target referent more likely than the competitor.

Fig. 6 displays the horizontal cursor position over time as a function of condition and groups. Looking at the time course of the decision process, there are clear temporal differences between conditions. In the RS group (upper panel), there are strong differences between all three conditions, with the VERB condition showing the earliest horizontal turn towards the target (the target is at $y = 1$) followed by OBJECT and by LEXICAL. We find a similar temporal pattern in participants' responses to reliable trials in the US group (lower panel), albeit with smaller differences between conditions. Nevertheless, descriptively, listeners in VERB trials turn to the target earliest, followed by OBJECT trials and LEXICAL trials.

3.3.2. Inferential assessment

We fitted Bayesian hierarchical linear models to turn-towards-target measurements as a function of CONDITION, GROUP, BLOCK, and their three-way interaction, using the Stan modeling language (Carpenter et al., 2016) and the package *brms* (Bürkner, 2016). The models included maximal random effect structures, allowing the predictors and their interactions to vary by participants (CONDITION and BLOCK) and experimental items (CONDITION, BLOCK,

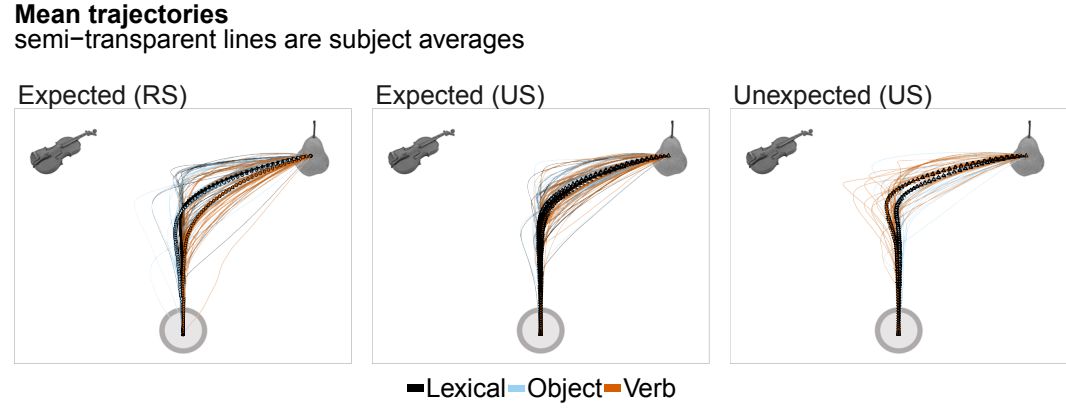

Fig. 5. Time and space normalized trajectories for all conditions in the reliable speaker group (left panel), for the expected form–function mappings in the unreliable speaker group (middle panel), and for the unexpected form–function mappings in the unreliable group (right panel). Semitransparent lines are averaged trajectories for individual participants.

and GROUP). We used weakly informative Gaussian priors centered around zero with $\sigma = 100$ for all population-level regression coefficients (e.g., Gelman, 2006), truncated Student-t priors centered at 0 (df = 3) for all standard deviations, and LKJ(2) priors for all correlation parameters. Four sampling chains with 2,000 iterations each were run, with a warm-up period of 1,000 iterations. For all relevant predictor levels and differences between them, we report 95% credible intervals (CIs)[3] and the posterior probability that a parameter β is smaller than zero $P(\beta < 0)$. The Bayesian inferential framework does not necessarily make dichotomous decisions (no significance threshold), but rather treats evidence as continuous. However, as a matter of fixing terminology, we judge there to be *compelling evidence* for an effect if zero is (by a reasonably clear margin) not included in the 95% CI and $P(\beta < 0)$ is close to zero or one.[4]

Fig. 7 displays the mean and 95% CIs of the posterior distribution (conditioned on the middle of the experiment, i.e., scaled block number = 0). Results are summarized in Tables 2 and 3. The LEXICAL trials serve as a baseline as listeners have to wait until the acoustic information about the referential expression becomes available. The acoustic information becomes available on average around 700 ms after the stimulus onset and it takes listeners around 910 ms to start turning towards the target (Reliable: $\beta = 904$, CI = (871; 940); Unreliable: $\beta = 920$, CI = (877;963)). In other words, mouse movements already indicate the integration of acoustic information around 210 ms after the relevant acoustic information becomes available. This resembles often cited temporal lags observed for eye-tracking studies which show that fixations lagged 200 ms behind relevant acoustic information (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Saslow, 1967; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

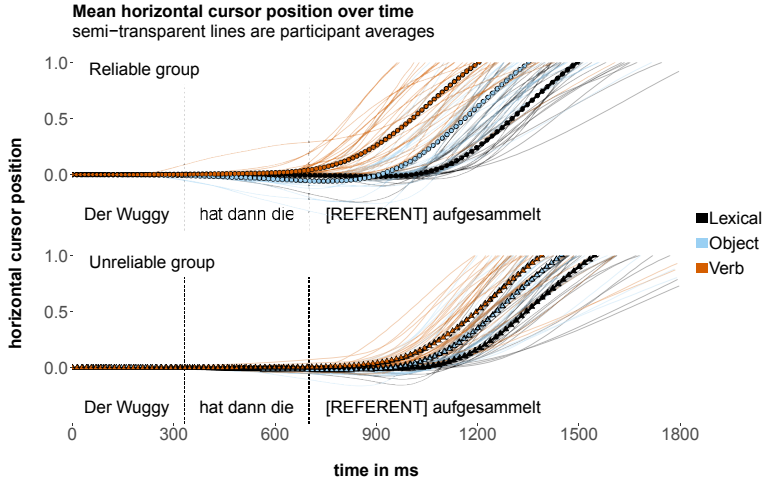

Fig. 6. Horizontal cursor position of space-normalized averaged trajectories in the reliable speaker group (top panel) and the unreliable speaker group (bottom panel) for Experiment 1. Semitransparent lines are averaged trajectories for individual participants.

There is compelling evidence that the intonationally informed conditions (verb and object) elicit earlier TTTs than the LEXICAL condition, with TTTs being earlier in the VERB condition (Reliable: β = 602, CI = (566;641); Unreliable: β = 739, CI = (694;780)) than in the OBJECT condition (Reliable: β = 768, CI = (737;797); Unreliable: β = 816, CI = (781;848)). The observed patterns suggest that listeners use intonational information prior to lexical information to anticipate the speaker's referential intention.

Looking across exposure groups, there is no compelling evidence that the LEXICAL disambiguation is affected by exposure to unreliable intonation (β = −16, CI = (−63;33), P (β < 0) = 0.75). This finding serves as a sanity check. The presence of unreliable intonation should not affect how listeners process lexical information. In contrast, there is compelling evidence that predictive interpretation of intonation is modulated by unreliable exposure. Both OBJECT trials (β = −47, CI = (−94;−4), $P(\beta < 0) = 0.98$) and VERB trials (β = −137, CI = (−191;−85), $P(\beta < 0) = 1$) exhibit overall later TTTs in the unreliable group.

These temporal effects changed dynamically across the course of the experiment (see Fig. 7B). There is compelling evidence that participants' anticipatory responses to OBJECT trials in the reliable group become quicker throughout the experiment, corresponding to a negative slope in Fig. 7B (β = −40, CI = (−60;−20), $P(\beta < 0) = 1$). There is also compelling evidence that participants' anticipatory responses to VERB trials in the unreliable group become slower throughout the experiment (β = 35, CI = (7;59), $P(\beta <; 0) = 0.01$).

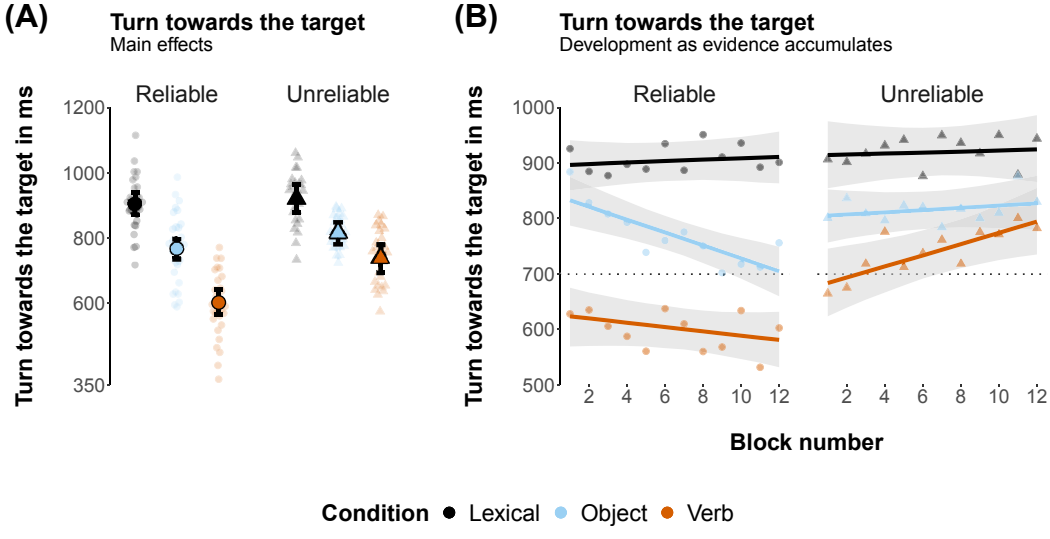

Fig. 7. Results of Experiment 1. (A) Posterior means and 95% credible intervals for the turn-towards-target measurement across conditions and groups (main effects). Semitransparent small points are average values for each participant. (B) Posterior means and 95% credible intervals for the turn-towards-target measurement across conditions and groups as a function of experimental blocks. Semitransparent points are average values. The dashed line indicates the acoustic onset of the referent.

Table 2
Posterior TTT estimates and 95% CIs for all three conditions in both listener groups

Parameter	Mean	95% CI
Object (Reliable)	768	(737;797)
Lexical (Reliable)	904	(871;940)
Verb (Reliable)	602	(566;641)
Object (Unreliable)	816	(781;848)
Lexical (Unreliable)	920	(877;963)
Verb (Unreliable)	739	(694;780)

Table 3
Posterior estimates and 95% CIs for TTT differences between conditions (rows 1–9) and posterior estimates and 95% CIs for the effect of experimental block for each condition (rows 10–15)

Parameter	Mean	95% CI	$P(\beta < 0)$
Object (Reliable)—Lexical (Reliable)	−135	(−168;−103)	1
Object (Reliable)—Verb (Reliable)	167	(132;200)	0
Lexical (Reliable)—Verb (Reliable)	302	(262;342)	0
Object (Unreliable)—Lexical (Unreliable)	−104	(−146;−65)	1
Object (Unreliable)—Verb (Unreliable)	77	(37;115)	0
Lexical (Unreliable)—Verb (Unreliable)	181	(133;234)	0
Object (Reliable)—Object (Unreliable)	−47	(−94;−4)	0.98
Lexical (Reliable)—Lexical (Unreliable)	−16	(−63;33)	0.75
Verb (Reliable)—Verb (Unreliable)	−137	(−191;−85)	1
Slope Object (Reliable)	−40	(−60;−20)	1
Slope Lexical (Reliable)	4	(−15;22)	0.32
Slope Verb (Reliable)	−13	(−36;10)	0.88
Slope Object (Unreliable)	7	(−16;28)	0.26
Slope Lexical (Unreliable)	3	(−22;30)	0.4
Slope Verb (Unreliable)	35	(7;59)	0.01

3.4. Discussion

The data of Experiment 1 confirm results from the previous literature, showing that intonational information, if used reliably according to the conventions of the respective speech community, can enable or facilitate the prediction of the meaning of a partially observed utterance (e.g., Dahan et al., 2002; Ito & Speer, 2008; Kurumada et al., 2014a; Roettger & Stoeber, 2017; Watson et al., 2008; Weber et al., 2006). The early pitch accent on the verb allows listeners to anticipate the intended referent long before the lexical material becomes available. Listeners also use the absence of an accent on the verb to anticipate the contrastive interpretation of the referent. This inference does not happen as fast as in the VERB condition but happens still substantially earlier than lexical disambiguation (LEXICAL > OBJECT > VERB).

Our data suggest that the predictive use of intonational cues depends on their estimated reliability. Listeners appear to weigh down the informational value of at least some

intonational cues in the unreliable group, but do not disregard them entirely, as these conditions still elicit anticipatory behavior in comparison to lexical disambiguation. Our data further suggest that incremental use of intonational cues changes dynamically throughout exposure. Exposure to reliable cues leads to earlier cue integration throughout the experiment in the OBJECT condition. Exposure to unreliable cues leads to later cue integration in the VERB condition. However, despite these dynamic changes and their differences between exposure groups, listeners predictively used the absence of a pitch accent on the verb already at the earliest stages of the experiment in both exposure groups. This suggests that predictive use of intonational cues can be differentially facilitated or modulated by exposure, but it is likely not just a mere task adaptation or experimental artifact. If it were, we would not expect listeners to start out with a predictive advantage at the beginning of the experiment. Instead, these results suggest that the predictive use of the absence of a pitch accent on the verb is part of the listeners' a priori assumptions about form–function mappings in their language.

A Bayesian model of the *evidential strength* of intonational cues qualitatively predicts the observed ordering relation of TTT measurement among conditions: VERB trials are predicted to be fastest because a pitch accent on the auxiliary is a strong cue to the upcoming discourse status of the referent. OBJECT trials are predicted to be slower because the absence of a pitch accent on the auxiliary is a weak cue, although they are predicted to be faster than lexical disambiguation. The model of dynamic adaptation to (partly) unreliable input from Section 2.4 further predicts that a flexible listener should (a) exhibit no noticeable decrease in TTT measurements throughout the experiment for VERB trials when exposed to only reliable mappings; (b) exhibit a facilitatory effect for OBJECT trials when exposed to only reliable mappings; and (c) exhibit an inhibitory effect of experience for VERB trials when exposed to occasionally unreliable mappings (see Fig. 4 in Section 3.2). All of these predictions are compatible with the data. Although the model does not make strong predictions for OBJECT trials in the US group, the outcome that was predicted to be most likely, namely no noticeable change in TTT measurements, is what the data indeed suggest.

In sum, the model's main categorical predictions about relative evidential strength and dynamic adaptation seem to be supported by the data. In particular, as predicted by gradual Bayesian belief updates, it seems that listeners, when confronted with partially unreliable input, do not stop exploiting intonational cues altogether; they rather seem to merely become more cautious. In this respect, the model makes even more specific predictions: We expect listeners to adapt differently to scenarios where only one cue is occasionally used unreliably, while the other is constantly reliable. The second experiment is designed to test these predictions.[5]

4. Experiment 2

The following experiment was preregistered on November 27, 2017, prior to data collection. The preregistration file can be retrieved from https://osf.io/49q2r/. All materials, raw data, and corresponding analysis scripts can be retrieved from https://osf.io/dnbuk/.

4.1. Method

Materials, procedure, and analyses were identical to Experiment 1, except for the distribution of stimuli. There were two exposure groups. The unreliable-VERB group was exposed to consistently natural OBJECT contours but occasionally encountered unreliable VERB contours; reversely, the unreliable-OBJECT group was exposed to consistently natural VERB contours but occasionally encountered unreliable OBJECT contours.

Participants were exposed to 12 blocks of 8 stimuli each. In the unreliable-VERB group, each block contained two reliable OBJECT trials, two reliable VERB trials, one unreliable VERB trial, and three LEXICAL trials. In the unreliable-OBJECT group, each block contained two reliable VERB trials, two reliable OBJECT trials, one unreliable OBJECT trial, and three LEXICAL trials.

Sixty native German speakers participated in the study, with 30 participants in each group. All participants had self-reported normal or corrected-to-normal vision and normal hearing (21 male, 39 female, M_{age} = 24.4 (SD = 3.4)). None of the participants had participated in the previous experiment.

4.2. Model predictions

Using the same procedure as in Experiment 1 (see Section 2.4 and Appendix A), we can derive the theoretical predictions for the temporal development of TTT measurements over the course of the experiment. A categorical summary of these predictions is shown in Fig. 8 (see also Fig. 4 for comparison). For each experimental condition, we tally how likely it is, from the model's point of view, that TTT measurements will increase, stay constant of decrease. The model predicts with a high degree of confidence that for the VERB condition in the speaker group that has only unreliable OBJECT trials (upper left panel of Fig. 8) there will be no noticeable difference in the TTT measure between the first and last block of the experiment. Intuitively, this is because the pitch accent on the auxiliary verb is already a strong cue and this will not be affected by occasional exposure to unreliable uses of OBJECT contours. In the OBJECT condition of the same group (lower left panel), predictions are less specific but nontrivial. The model almost completely rules out that TTT measurements should increase, and it considers constancy slightly more likely than a decrease. In intuitive terms, since listeners are exposed to more reliable instances of the OBJECT contour, indeed twice as many as unreliable ones, the model considers it possible that the evidential value of an absent pitch accent could stay relatively constant, if not even rise slightly over the course of the experiment.

For the speaker group that has only unreliable VERB trials, we predict that VERB trials should almost certainly not become faster. The model considers it possible that TTT become slower, but still considers it most likely that, given the initially high evidential value of the pitch accent and the fact that only one out of three occurrences of VERB are unreliable, TTT measurements do not change noticeably over the course of

the experiment. For OBJECT trials in that group, we predict a high probability of trials becoming faster over the course of experiment, that is, a decrease in TTT measurement. This is because the OBJECT contour is reliably used throughout the experiment and so the model predicts that this initially rather weak cue is strengthened during the course of the experiment, regardless of the occasionally occurring unreliable VERB trials.

4.3. Results

Following preregistered protocols, the whole dataset of participants was excluded whenever they (a) exhibited more than 10% errors, or (b) exhibited movement behavior violating instructions in more than 15% of the trials, or (c) exhibited initiation times above 350 ms in more than 15% of the trials. For each exclusion criteria, we had to exclude one participant. Additionally, we excluded two participants due to a malfunctioning of the experimental software. (This exclusion was not anticipated in the preregistration.) Trials with initiation times greater than 350 ms (1.5%) and incorrect responses (1.8%) were discarded on a trial-by-trial basis. Trials that exhibited movement behavior violating instructions were discarded, too (1.5%).

Fig. 9 displays averaged trajectories. Looking at the time course of the decision process (Fig. 9A), there are again clear temporal differences between conditions in both groups. In the unreliable-OBJECT group (top panel of Fig. 9A), there are strong differences between LEXICAL and OBJECT trials on one hand and VERB trials on the other. In the unreliable-VERB group (bottom panel), this VERB advantage is noticeably smaller.

Fig. 10 and Tables 4 and 5 summarize the statistical results, following the same presentation format as for Experiment 1. Results for LEXICAL disambiguation, serving as a baseline, replicate the observed patterns from Experiment 1. Listeners turn towards the target around 900 ms after the acoustic onset of the stimuli (Unreliable object: $\beta = 884$, CI = (849;921); Unreliable verb: $\beta = 909$, CI = (868;949)). Moreover, there is compelling evidence that

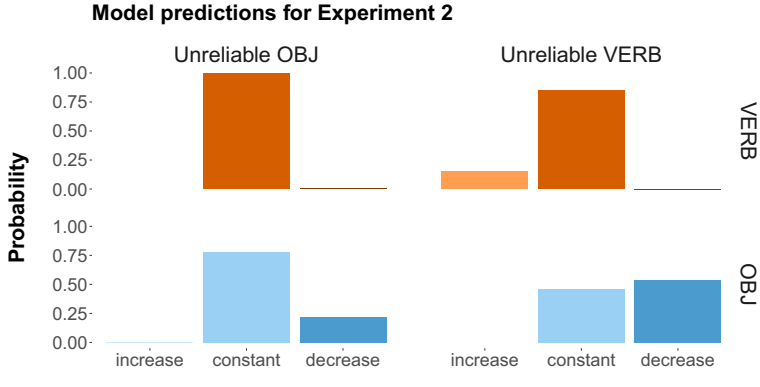

Fig. 8. Categorical predictions about the temporal development of the TTT measure throughout Experiment 2.

early intonational cues from the VERB and OBJECT conditions indeed elicited earlier TTTs than the LEXICAL condition, with TTTs being earlier in the VERB condition (Unreliable object: $\beta = 576$, CI = (534;623); Unreliable verb: $\beta = 716$, CI = (656;772)) than in the OBJECT condition (Unreliable object: $\beta = 828$, CI = (791;864); Unreliable verb: $\beta = 792$, CI = (751;830)).

Looking across groups, neither OBJECT nor LEXICAL shows clear indications of an impact of the group manipulation (Object: $\beta = 35$, CI = $(-16;83)$, $P(\beta < 0) = 0.09$; Lexical: $\beta = -25$, CI = $(-73;28)$, $P(\beta < 0) = 0.84$). VERB trials, however, are reacted to slower in the unreliable-VERB group ($\beta = -140$, CI = $(-208;-70)$, $P(\beta < 0) = 1$).

Fig. 10B displays how TTT measurements change over the experiment. In comparison to the patterns described in Experiment 1, there is no compelling evidence that the present effects change throughout the experiment. According to our preset terminological convention, we cannot claim that there is compelling evidence that the development of participants' anticipatory behavior over the course of the experiment (slope of the lines) is different from zero (= a flat line) (see Table 5), although there is "weak evidence" that the slope of the VERB condition in the unreliable-VERB group is positive ($\beta = 22$, CI = $(-9;54)$, $P(\beta < 0) = 0.08$).

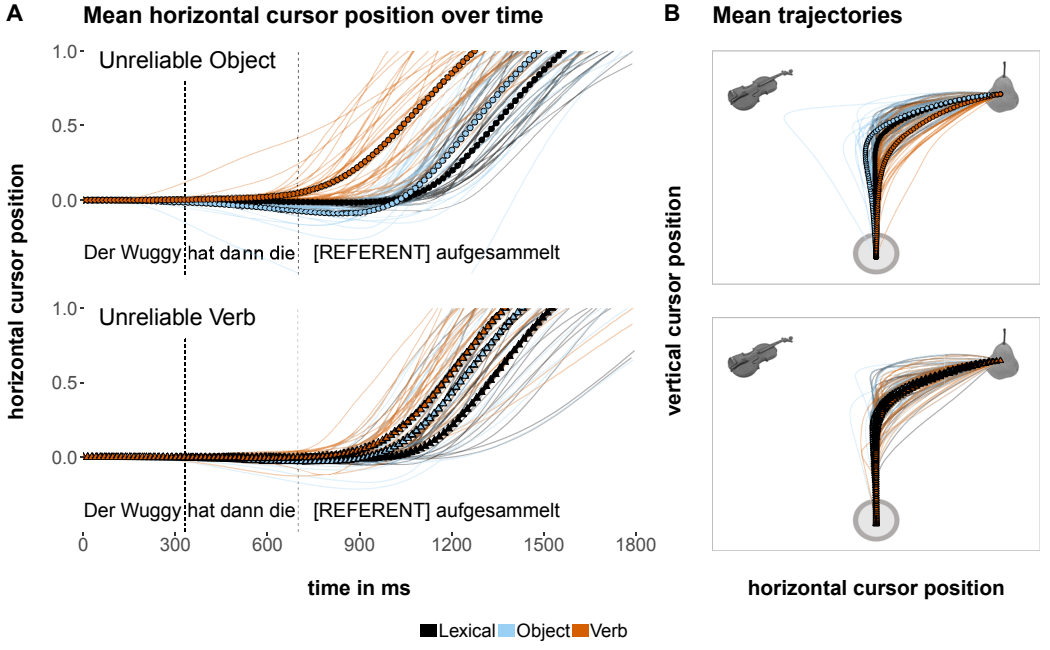

Fig. 9. Trajectories of Experiment 2. (A) Horizontal cursor position of space-normalized averaged trajectories in the unreliable-OBJECT group (top) and the unreliable-VERB group (bottom). (B) Time- and space-normalized averaged trajectories for both groups. Semitransparent lines are averaged trajectories for individual participants.

4.4. Discussion

The observed data are largely compatible with the model's predictions. No slope coefficient in the regression analysis is credibly different from zero, suggesting that all TTT measurements were mostly constant over the course of the experiment. This is the model's most likely prediction for all conditions except for the OBJECT condition in the unreliable-VERB group, where the model would have predicted a likely decrease in TTT measurement (see Fig. 8).

Despite only weak evidence for dynamic changes of TTT measurements throughout the experiment, there are suggestive patterns when we compare the beginning and end of the experiment. In the unreliable-VERB group, OBJECT trials are initially similarly slow as LEXICAL trials (CI overlap, see Fig. 10). LEXICAL trials seem to become slower and OBJECT trials faster, leading to a substantial differences between these conditions by the end of the experiment. Thus, listeners seem to learn to use the absence of a pitch accent on the auxiliary as a predictive cue to an upcoming contrastive referent. Learning happens despite occasional unreliable form-function mappings in the VERB condition.

Contrary to this, at the beginning of the experiment, the VERB condition in the unreliable-VERB group starts with a temporal advantage over OBJECT. However, throughout the experiment, VERB appears to become slower, approaching the temporal performance of

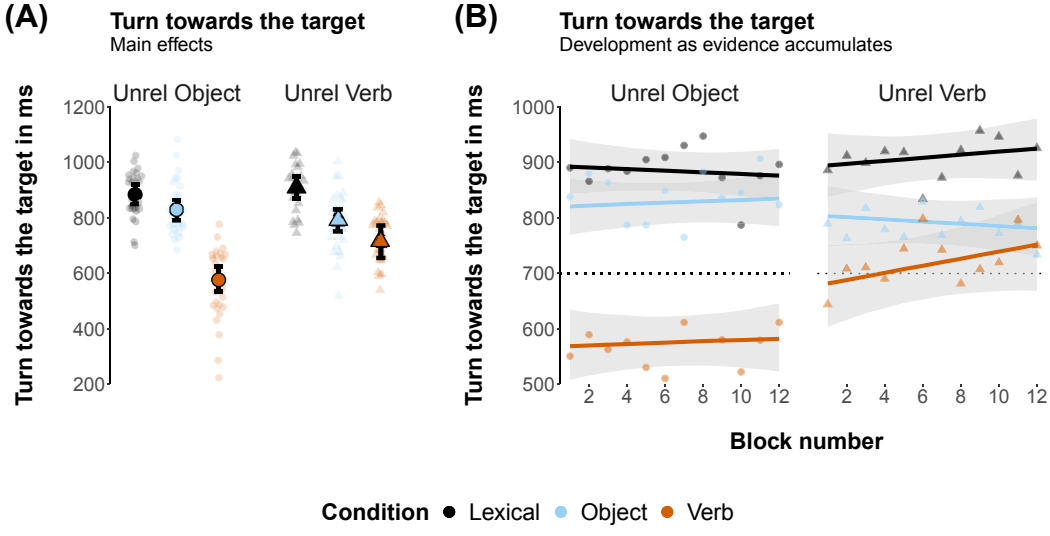

Fig. 10. Results of Experiment 2. (A) Posterior means and 95% credible intervals for the turn-towards-target measurement across conditions and groups (main effects). Semitransparent small points are average values for each participant. (B) Posterior means and 95% credible intervals for the turn-towards-target measurement across conditions and groups as developing through the experiment. Semitransparent points are overall average values. Dashed line indicates the acoustic onset of the referent.

Table 4
Posterior TTT estimates and 95% CIs for all three conditions in both listener groups

Parameter	Mean	95% CI
Object (Unrel. Object)	828	(791;864)
Lexical (Unrel. Object)	884	(849;921)
Verb (Unrel. Object)	576	(534;623)
Object (Unrel. Verb)	792	(751;830)
Lexical (Unrel. Verb)	909	(868;949)
Verb (Unrel. Verb)	716	(656;772)

Table 5
Posterior estimates and 95% CIs for TTT differences between conditions (rows 1–9) and for the effect of experimental block for each condition (rows 10–15)

Parameter	Mean	95% CI	$P(\beta < 0)$
Object (Unrel. Object)—Lexical (Unrel. Object)	−57	(−89;−26)	1
Object (Unrel. Object)—Verb (Unrel. Object)	252	(204;300)	0
Lexical (Unrel. Object)—Verb (Unrel. Object)	308	(262;355)	0
Object (Unrel. Verb)—Lexical (Unrel. Verb)	−117	(−154;−82)	1
Object (Unrel. Verb)—Verb (Unrel. Verb)	76	(20;135)	0
Lexical (Unrel. Verb)—Verb (Unrel. Verb)	194	(134;249)	0
Object (Unrel. Object)—Object (Unrel. Verb)	35	(−16;83)	0.09
Lexical (Unrel. Object)—Lexical (Unrel. Verb)	−25	(−73;28)	0.84
Verb (Unrel. Object)—Verb (Unrel. Verb)	−140	(−208;−70)	1
Slope Object (Unrel. Object)	4	(−18;24)	0.34
Slope Lexical (Unrel. Object)	−5	(−25;16)	0.69
Slope Verb (Unrel. Object)	4	(−21;31)	0.39
Slope Object (Unrel. Verb)	−7	(−31;17)	0.7
Slope Lexical (Unrel. Verb)	9	(−15;34)	0.23
Slope Verb (Unrel. Verb)	22	(−9;54)	0.08

OBJECT by the end of the experiment (CI are heavily overlapping). Thus, there are some suggestions that listeners appear to selectively unlearn the expected speaker production probabilities for intonation contours with a pitch accent on the verb and learn to predictively use the form–function mapping in the OBJECT condition.

In conclusion, the model's predictions seem to be generally compatible with the data. A suggestive trend in the data towards increasing TTTs for VERB trials in the unreliable-VERB group might be further taken as support for an increasing trend predicted by the model. Overall, the data still seem to support the general prediction that unreliable occurrences of one type of intonational cue do not necessarily lead to a neglect of other types of cues. This is clearly seen in the results shown in Fig. 8A where the VERB condition was still very fast despite occasional unreliable uses of OBJECT contours. Similarly, the OBJECT condition was still faster than lexical disambiguation even when unreliable VERB contours were observed.

5. General discussion

The presented experiments investigated how listeners use information from intonation to predictively anticipate an utterance's eventual meaning and how they adapt their anticipatory behavior in light of reliable or unreliable uses of intonational cues. Using mouse tracking, this study replicates earlier findings showing that listeners can, in principle, rapidly integrate intonational cues to predict the speaker's intended meaning. When listeners hear an early pitch accent (or the absence thereof), their manual response dynamics indicate an early bias towards one interpretation over another (Roettger & Stoeber, 2017). This is in line with a large body of evidence that listeners use intonational information to predict referential intentions (e.g., Dahan et al., 2002; Ito & Speer, 2008; Kurumada et al., 2014a; Watson et al., 2008; Weber et al., 2006).

We also showed that, over the course of the experiment, consistent reliable input leads to earlier anticipation of initially weak intonational cues (i.e., the absence of a pitch accent on the early verb), while unreliable input delayed anticipation based on an initially strong intonational cue (i.e., the presence of an early pitch accent on the verb).

To account for both the time course of information integration and how it changes over the course of the experiment, we introduced a Bayesian model of predictive cue integration. Much previous work has assumed that a listener's interpretation can be modeled as Bayes rule based on the listener's assumptions about the speaker's utterances production behavior (e.g., Frank & Goodman, 2012; Franke & Jäger, 2016; Kehler & Rohde, 2015; Russell, 2012). The model presented here extends this line of work in several directions. First, it addresses aspects of online interpretation (e.g., Werning & Cosentino, 2017). Second, we suggest a concrete linking hypothesis, relating a quantitative notion of evidential strength of a cue to aspects of mouse trajectories in a two-alternative interpretation task. Third, the paper formulates a theory of observation-based belief updates in terms of hierarchically structured listener beliefs. It is shown how such a model, though complex, can be simplified to a very convenient formulation in terms of increments of non-normalized weights, when combined with the aforementioned notion of evidential strength. Our model predicts interesting asymmetries in listeners' responses and their temporal development, which are generally supported by the data.

5.1. Possible limitations of this study

The stimuli in our experiments have been resynthesized in such a way that the pitch modulations arguably sound slightly less natural (see stimuli in our repository). We generated all stimuli from the same base contour (the VERB contour) and changed the intensity envelope and the (f_0) contour in order to create three intonation categories. That way we had complete control over the temporal structure of stimuli, resulting in sentences in which the lexical acoustic landmarks become available at the same time across conditions (e.g., the acoustic onset of "pear" starts at the same time across conditions). It further

reduced the available cues for listeners to those that we are interested in (the pitch modulation). More natural stimuli exhibit acoustic differences between conditions distributed throughout the entire utterance, making a clear interpretation of the temporal integration of specific cues difficult. However, we do not think that our results can be explained by the potential unnaturalness of our stimuli or auditory artifacts due to the following arguments:

First, there is ample evidence that speech adaptation mechanisms are very similar across different levels of distortion. Listeners adapt to distorted speech patterns in production and perception. Speakers rapidly adapt their speech production strategies to a variety of auditory perturbations (e.g., Jones & Munhall, 2000; Lane & Tranel, 1971) and listeners exposed to noise-vocoded sentences are able to tune into these patterns within short periods of time (e.g., Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). Second, our stimuli are identical across groups, so any confounds due to the acoustic makeup of the stimuli do not selectively affect the unreliable group and thus do not explain why conditions elicit different patterns across exposure groups.

In sum, although we do not believe that our results are affected by artifacts introduced by the resynthesis procedure, our findings should be replicated with more naturally sounding stimuli. In fact, as part of a follow-up study, we replicated the results of the reliable group with more natural stimuli and a different source speaker. Both the general differences between conditions and the dynamic adaptation effects qualitatively replicate (Roettger, Franke, & Cole, 2019). The data are freely available here: https://osf.io/xf8be/.

5.2. Possible extensions of the model

We maintain that the model presented here is a good step forwards in our effort to understand how listeners might form rational predictions based on early intonational cues and how they might update their predictive processing strategies in light of observations of speaker usage. The model we presented here is, by virtue of being a model, an idealized simplification. Our starting point is a rational analysis of the to-be-explained behavior (Anderson, 1990; Chater & Oaksford, 1999). In this tradition, formal modeling serves the purpose of highlighting discrepancies between reality and an idealized picture. However, already at a conceptual level, some of the idealizations inherent in our model are worthy of criticism.

For example, we have assumed an ideal observer who does not make any mistakes in perceiving a given contour and does not forget any previously observed speaker action. Although the inherent variability in intonational categories ultimately needs to be related to variability observed in other domains of speech, the proposed calibration might be much closer related to adjustments to variability in the choice of other linguistic phenomena, such as the choice of overinformative modifiers (Grodner & Sedivy, 2011). The gradient acoustic realization of postulated intonational categories shows highly overlapping distributions of acoustic parameters, notoriously making both the automatic retrieval as well as annotator agreement difficult (e.g., Cole & Shattuck-Hufnagel, 2016, for a recent discussion). Thus, intonational categories have to be inferred from gradient distributions and are likely to be perceived with some stochastic noise. Combining the model

presented here with a component of noise-perturbed perception (e.g., Kleinschmidt & Jaeger, 2015) is a plausible and possibly rewarding extension for future work.

In our model, we further posited what we referred to as plausible constraints on the production likelihood of producing pitch accent on the verb or not. These constraints are plausible in the sense that they agree with our native intuitions about German and they are in line with recent production studies. However, a more principled quantification of production likelihoods would be desirable for future iterations. Corpus studies or gating experiments could yield more precise estimates of how a partial utterance with particular intonational cues is likely continued in a given discourse context.

Future work should further investigate several other aspects which the present model has not addressed. For one, we have only targeted a population-level aggregate, leaving potential individual differences among participants out of the picture. This may leave interesting interindividual differences unnoticed and might even blur the general picture (Estes, 1956; Estes & Maddox, 2005). For another, we have only focused on the turn-towards-target measure which we have linked to the dynamically developing posterior beliefs of listeners during the process of observing an utterance. But this is only one aspect of the rich data associated with mouse trajectories. We might speculate, for instance, that higher-order uncertainty—as represented by lower or higher Dirichlet weights in the hierarchical listener belief model—might also be reflected in other aspects of the trajectories, such as momentary speed (Dotan et al., 2018) or fluctuation patterns (Dale, Roche, Snyder, & McCall, 2008).

Finally, it is not difficult to conceive of several plausible alternatives to the picture of listener adaptation based on Bayesian belief revision explored here. There is even some indication in our data that an alternative explanation might be needed. When looking at the temporal development of TTT measurements in all experimental groups which encountered unreliable VERB conditions (Fig. 7B and Fig. 10B), we see that already after the first block there seems to be a large effect of the unreliability of VERB trials. The model presented here does not capture this. After the first block, listeners have only observed one unreliable instance of a VERB contour. The model presented here would only predict a mild change as compared to the interpretation where all VERB contours were perceived to be used reliably. It is conceivable if not likely that listeners have a more elaborate belief update process than modeled here. Already after the first example of an unreliable use of what is normally a high-fidelity cue, listeners might be immediately alerted. At least two routes could be explored in future modeling:

First, upon observing already the first obvious violation of expectations, listeners might adjust their readiness to deviate from their default beliefs, for example, modeled here by adjusting the ω parameter in the higher-order beliefs. Second, we might extend the hierarchical modeling altogether to include listeners' beliefs about different speaker types. After all, listeners' expectations about speaker behavior might not be an undifferentiated whole, but allow for tracking individual speakers simultaneously (e.g., Grodner & Sedivy, 2011; Kleinschmidt & Jaeger, 2015; Kraljic & Samuel, 2007; Pogue, Kurumada, & Tanenhaus, 2016; Trude & Brown-Schmidt, 2012). One speaker type might be a "weird speaker" and observing a single highly unexpected utterance may be enough to shift probability mass

quickly to the belief, which initially has a very low probability, that the interlocutor in front of us is of an abnormal variety. These considerations suggest that further empirical research and modeling efforts should be dedicated specifically to the first instances of listener adaptation, and to investigate how a listener who has adapted to one speaker would confront another.

Though idealized and exploratory, we nevertheless consider this modeling work a helpful step forward in understanding predictive processing of intonational cues. The model presented here is a proof of concept that intonational processing can be described as rational integration of observational evidence and that dynamic adaptation to a speaker's possibly idiosyncratic speech behavior can be explained as rational belief update of a prior belief that captures assumptions about how speakers of a language normally behave.

6. Conclusion

Intonation provides important early cues to speaker-intended information. We presented data that suggest that listeners are able to rapidly integrate intonational cues during online processing. As demonstrated by the time at which participants started to move their mouse consistently towards the target referent, listeners picked up on the presence or absence of a pitch accent on an early auxiliary verb as a predictive cue about whether the speaker will likely refer to a discourse-mentioned or discourse-new referent. By manipulating the reliability of intonational cues in a between-subject design, we further found that listeners generally seem to anticipate a reliable mapping and to start to predictively use intonational information early on. This suggests that rapid intonational cue integration is not just a rational adaptation to the experimental task, but a general predisposition of language users to use intonation in accordance with their language. Over the course of the experiment consistent reliable input leads to earlier turns towards the target of initially weak intonational cues (the OBJECT condition with the absence of a pitch accent on the verb), while partly unreliable input impeded predictive use of an initially strong intonational cue (the VERB condition with a pitch accent on the verb).

We presented a novel and exploratory model of rational incremental belief update and belief dynamics to argue that these qualitative patterns observed in our experimental data are compatible with the idea that listeners rationally and rapidly integrate intonational information and update their expectations about speaker production likelihoods dynamically. We ran a second experiment to test these model predictions and were able to substantiate the models qualitative predictions on a new dataset. To conclude, this study contributes to our knowledge about how listeners deal with the omnipresent variability in mapping parts of the speech signal onto communicative functions. Listeners rapidly integrate bottom-up acoustic information and weigh it against their top-down expectations about likely prosodic patterns. Moreover, listeners are very rational when they evaluate the reliability, that is, the usefulness of parts of the signal to predict what the speaker intends to communicate.

Acknowledgments

Timo Roettger's work was supported by the "Zukunftskonzept" of the University of Cologne as part of the Excellence Initiative. Michael Franke's work was supported by the Priority Program XPrag.de (DFG Schwerpunktprogramm 1727). We would like to thank Nastassja Bremer and Kim Rimland for their help during data collection. We are grateful for Limor Raviv's valuable feedback on an earlier draft of this manuscript and we would like to thank three anonymous reviewers for their insightful comments and suggestions. All remaining errors are our own.

Notes

1. Here, "heading towards the target" is operationalized by approximating the first derivative to the x- and y-coordinates of a trajectory; see function `get_TTT_derivative()` in our analysis scripts.
2. We also preregistered, measured, and analyzed reaction times and two spatial parameters, but focus here on said temporal online measurement. See our online repository for more details.
3. A 95% credible interval demarcates the range of values that comprises 95% of probability density or mass of our posterior beliefs (e.g., Jaynes, 2003; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016).
4. Note that we preregistered an analysis within the null hypothesis significance testing framework. Due to severe convergence issues with complex random effect structures, we were not able to run the desired regression models. Simpler models converged and provided comparable results to the presented Bayesian analysis. However, because the exclusion of particular random slopes can increase the Type-I error rate (in a frequentist framework) and underestimate the group-level variability, we decided to back up the preregistered analysis with the conceptually desired random effect structure in the present Bayesian analysis. This approach resulted in the same overall results and can be considered more conservative. Both analyses and their results can be assessed via the R scripts in our OSF repository.
5. We note here that the model proposed in Section 2.4 and Appendix A was conceived after the analysis of the first experiment in a post hoc fashion. Based on the model, we then generated new hypotheses that we tested in the second experiment.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Baumann, S., & Grice, M. (2006). The intonation of accessibility. *Journal of Pragmatics*, *38*(10), 1636–1657.
Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology*, *3*, 255–309.
Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer. [computer program]. version 6.0.17.
Bolinger, D. (1972). Accent is predictable (if you're a mind-reader). *Language*, 633–644.
Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.
Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE*, *5*(5), e10773.
Bürkner, P.-C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.
Buxo-Lugo, A. F. (2017). Communicative context, expectations, and adaptation in prosodic production and comprehension. Doctoral dissertation, University of Illinois at Urbana-Champaign.
Buxó-Lugo, A., & Watson, D. G. (2016). Evidence for the influence of syntax on prosodic parsing. *Journal of Memory and Language*, *90*, 1–13.
Calhoun, S. (2007). Information structure and the prosodic structure of English: A probabilistic relationship. Doctoral dissertation, University of Edinburgh.
Cangemi, F., Krüger, M., & Grice, M. (2015). Listener-specific perception of speaker-specific production in intonation. In S. Fuchs, D. Pape, C. Petrone, & P. Perrier *Individual differences in speech production and perception*, (pp. 123–145). Cambridge, MA: Peter Lang Publishing Group.
Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*, 1–37.
Chater, N., & Oaksford, M. (1999). Ten year of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.
Chodroff, E., & Cole, J. (2018). Information structure, affect, and prenuclear prominence in American English. *Proceedings of the annual conference of the international speech communication association, interspeech* (Vol. 2018, pp. 1848–1852). ISCA.
Clopper, C. G., & Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, *39*(2), 237–245.
Cole, J., & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*(1).
Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, *77*(6), 2142–2156.
Cruttenden, A. (1997). *Intonation*. Cambridge: Cambridge University Press.
Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(2), 141–201.
Dahan, D. (2015). Prosody and language comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(5), 441–452.
Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292–314.
Dale, R., Roche, J., Snyder, K., & McCall, R. (2008). Exploring action dynamics as an index of paired-associate learning. *PLoS ONE*, *3*(3), e1728.
Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241.
Dotan, D., Meyniel, F., & Dehaene, S. (2018). On-line confidence monitoring during decision making. *Cognition*, *171*, 112–121.
Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53* (2), 134–140.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*(3), 403–408.

Féry, C., & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics*, *36*(4), 680–703.

Fischer, M. H., & Hartmann, M. (2014). Pushing forward in embodied cognition: May we mouse the mathematical mind? *Frontiers in Psychology*, *5*, 1315.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336* (6084), 998.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44.

Freeman, J. B., & Ambady, N. (2010). Mousetracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America*, *27*(4), 765–768.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Grice, M. (1995). *The intonation of interrogation in Palermo Italian: Implications for intonation theory*. Berlin, Germany: de Gruyter.

Grice, M., Lohnstein, H., Röhr, C. T., Baumann, S., & Dewald, A. (2012). The intonation of verum focus and lexical contrast. Proceedings of *Phonetik & Phonologie 8*, October 12, 2012 (pp. 22–23). Jena: Friedrich-Schiller-Universität.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics*, *64*, 90–107.

Grodner, D. J., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. In E. Gibson, & N. J. Pearlmutter (Eds.), *The processing and acquisition of reference* (pp. 239–272). MIT Press.

Gussenhoven, C. (1984). On the grammar and semantics of sentence accents. Doctoral dissertation, Radboud University Nijmegen.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge, UK: Cambridge University Press.

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, *18*(3), 384–401.

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, *36*(1–2), 31–43.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*(2), 541–573.

Ito, K., Speer, S. R., & Beckman, M. E. (2004). Informational status and pitch accent distribution in spontaneous dialogues in English. In B. Bel & I. Marlien (Eds.), Proceedings of Speech Prosody (pp. 279–282). Nara, Japan: SProSIG.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127* (1), 57–83.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Jaynes, E. T., & Kempthorne, O. (1976). Confidence intervals vs. Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 6b, pp. 175–257). The University of Western Ontario Series in Philosophy of Science. Dordrecht, the Netherlands: Springer.

Jeffrey, R. (2002). *Subjective probability: The real thing*. Princeton, NJ: Princeton University Press.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of f0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America*, *108*(3), 1246–1251.

Jun, S.-A. (2007). *Prosodic typology: The phonology of intonation and phrasing*. Oxford, UK: Oxford University Press.

Jun, S.-A.(2014). *Prosodic typology II: The phonology of intonation and phrasing*. Oxford, UK: Oxford University Press.

Kehler, A., & Rohde, H. (2015). Pronominal reference and pragmatic enrichment: A Bayesian account. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1063–1068). Austin, TX: Cognitive Science Society.

Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*, 1–16.

Kieslich, P. J., Schoemann, M., Grage, T., Hepp, J., & Scherbaum, S. (2019). Design factors in mousetracking: What makes a difference? *Behavior Research Methods*.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014a). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014b). Rapid adaptation in online pragmatic interpretation of contrastive prosody. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36 th Annual Meeting of the Cognitive Science Society* (pp. 791–796). Austin, TX: Cognitive Science Society.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2017). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin & Review*, *25*(3), 1153–1160.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge: Cambridge University Press.

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research*, *14*(4), 677–709.

Magnuson, J. S. (2005). Moving hand reveals dynamics of thought. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 9995–9996.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Peppé, S., Maxim, J., & Wells, B. (2000). Prosodic variation in southern British English. *Language and Speech*, *43*(3), 309–334.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. (Doctoral dissertation). Cambridge, MA: MIT.

Pierrehumbert, J., & Hirschberg, J. B. (1990). The meaning of intonational contours in the interpretation of discourse. *Intentions in Communication*, 271–311.

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under-and over-informative prenominal adjective use. *Frontiers in Psychology*, *6*, 2035.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Roettger, T. B. (2017). *Tonal placement in Tashlhiyt: How an intonation system accommodates to adverse phonological environments*. Berlin, Germany: Language Science Press.

Roettger, T. B., Franke, M., & Cole, J. (2019). Testing the relevance of prenuclear accents for predicting intonational meaning in German. In *19th International Congress on Phonetic Sciences (ICPhS)*, Melbourne, Australia.

Roettger, T. B., Mahrt, T., & Cole, J. (2019). Mapping prosody onto meaning–the case of information structure in American English. *Language, Cognition and Neuroscience*, 1–20.

Roettger, T. B., & Stoeber, M. (2017). Manual response dynamics reflect rapid integration of intonational information during reference resolution. In G. Gunzelmannn, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 3010–3015). Austin, TX: Cognitive Science Society.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, *1*(1), 75–116.

Russell, B. (2012). Probabilistic reasoning and the computation of scalar implicatures. Doctoral dissertation, Brown University.

Saslow, M. (1967). Effects of components of displacement-step stimuli upon latency for saccadic eye movement. *JOSA*, *57*(8), 1024–1029.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tomlinson, J., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language*, *69*(1), 18–35.

Tomlinson, J., Gotzner, N., & Bott, L. (2017). Intonation and pragmatic enrichment: How intonation constrains ad hoc scalar inferences. *Language and Speech*, *60*(2), 200–223.

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, *27*(7–8), 979–1001.

Turco, G., Dimroth, C., & Braun, B. (2013). Intonational means to mark verum focus in German and French. *Language and Speech*, *56*(4), 461–491.

Turnbull, R. (2017). The role of predictability in intonational variability. *Language and Speech*, *60*(1), 123–153.

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L+H*. *Cognitive Science*, *32*(7), 1232–1244.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. *Language and Speech*, *49*(3), 367–392.

Werning, M., & Cosentino, E. (2017). The interaction of Bayesian pragmatics and lexical semantics in linguistic interpretation: Using event-related potentials to investigate hearers' probabilistic predictions. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 3504–3509). Austin, TX: Cognitive Science Society.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143.

Appendix: A hierarchical model of listener beliefs

The four goals of this section are (a) to spell out a Dirichlet-multinomial model of the listener's beliefs about the speaker's probabilistic choice behavior, (b) to show that this model conservatively extends the results from Section 2.3 on the relative evidential strength of the relevant partial utterances, (c) to describe how learning

from observation is captured in this model, and (d) to show how predictions can be derived regarding the temporal development of TTT measurements during the course of the experiment. Let the speaker's probabilistic production behavior be given by a conditional probability distribution, here written as a row-stochastic matrix S, with S_{ik} the probability of producing utterance k for meaning i. The listener's higher-order beliefs $P(S)$ assign a probability to each speaker matrix S. A compact and convenient way of defining such a higher-order probability function is to use a Dirichlet distribution. A Dirichlet distribution is a probability distribution over the class of all discrete probability distributions over some finite set of events (a so-called multinomial distribution)—in our case, the set of possible utterances. The Dirichlet distribution takes as parameter a vector of positive numbers of so-called Dirichlet weights. Let L be a matrix of the same dimensions as S such that each row L_i is the vector of Dirichlet weights for probability vector S_i. We assume that $P(S \mid L) = \prod_i Dirichlet(S_i \mid L_i)$, where by definition $\text{Dirichlet}(S_i \mid L_i) \propto \prod_k (S_{ik})^{L_{ik}-1}$. An example of Dirichlet weights is given in the table in Fig. 2A. It is important to notice that Dirichlet weights do not need to sum up to one. Rather, the absolute sum of Dirichlet weights encodes the confidence of higher-order beliefs: the higher $\sum_k L_{ik}$, the more certain the listener is that the speaker's production behavior S_i is close to the mean expectation of the corresponding Dirichlet distribution, which is easy to calculate as: $\int \text{Dirichlet}(S_i \mid L_i)\ S_{ik}\ \mathrm{d}S_i = \frac{L_{ik}}{\sum_l L_{il}}$.

If the listener has uncertainty about the speaker's production likelihoods, we must integrate over this higher-order uncertainty to compute the posterior odds. But for a Dirichlet-multinomial model we can give the likelihood ratio simply in terms of normalized Dirichlet weights:

$$\begin{aligned}\frac{P(r_i \mid u_k)}{P(r_j \mid u_k)} &= \int\int \frac{P(r_i)}{P(r_j)}\ \frac{\text{Dirichlet}(S_i \mid L_i)}{\text{Dirichlet}(S_j \mid L_j)}\ \frac{S_{ik}}{S_{jk}}\ \mathrm{d}S_i \mathrm{d}S_j \\ &= \frac{P(r_i)}{P(r_j)}\ \frac{\int \text{Dirichlet}(S_i \mid L_i)\ S_{ik}\ \mathrm{d}S_i}{\int \text{Dirichlet}(S_j \mid L_j)\ S_{jk}\ \mathrm{d}S_j} \\ &= \underbrace{\frac{P(r_i)}{P(r_j)}}_{\text{prior odds}}\ \underbrace{\frac{L_{ik}/\sum_l L_{il}}{L_{jk}/\sum_l L_{jl}}}_{\text{likelihood ratio}}\end{aligned}$$

This result implies that, in order to calculate the evidential strength of a (partial) utterance, we just need to normalize the matrix of Dirichlet weights and proceed as before.

The Dirichlet distribution is the conjugate prior of the multinomial distribution. This allows for a very simple formulation and implementation of the listener's belief updates. If the listener's beliefs are given by L prior to observing an utterance u_k for meaning r_i, then they will be L' after this observation, with L' exactly like L except that $L'_{ik} = L_{ik} + 1$. To see this, Let $\vec{x}$ be a vector with x_k the number of observations of u_k for meaning r_i. Then, the listener's posterior beliefs about S_i after observing $\vec{x}$ are:

$$
\begin{aligned}
P(S_i \mid \vec{x}) &\propto \text{Dirichlet}(S_i \mid L_i)\ \text{Multinomial}(\vec{x} \mid S_i) \\
P(S_i \mid \vec{x}) &\propto \prod_k (S_{ik})^{L_{ik}-1} \prod_k (S_{ik})^{x_k} \;=\; \prod_k (S_{ik})^{L_{ik}+x_k-1} \\
P(S_i \mid \vec{x}) &= \text{Dirichlet}(S_i \mid L_i'), \quad \text{where } L_{ik}' = L_{ik} + x_k.
\end{aligned}
$$

To derive predictions about changes of the TTT measure over the course of the experiment, define a function $F_{G,u}$, for each experimental group G and utterance u. This function maps each triple of parameters p_V, ε_V, and ω, which specify a concrete matrix of Dirichlet weights (see Section 2.4), to the difference between the TTT measure after the last and the first block of the experiment which we would expect for the given parameter triple. For example, for the RS group and the VERB condition, we obtain:

$$
\begin{aligned}
F_{\text{RS,V}}(p_V, \epsilon_V, \omega) &= \text{expected TTT after block 12} - \text{expected TTT after block 1} \\
&= \exp\left(1 - \frac{p_V + 24/\omega}{\epsilon_V}\right) - \exp\left(1 - \frac{p_V + 2/\omega}{\epsilon_V}\right)
\end{aligned}
$$

We do not commit to a single triple of parameters here, but rather maintain a (modeler's) prior distribution $P(p_V, \varepsilon_V, \omega)$ over them (see Section 2.4). Since the assumed link function is only approximate, the values of $F_{G,u}$ should only be interpreted as indications of the categorical trend we predict. Positive values of $F_{G,u}$ imply an increase in TTT measurements, negative values a decrease, and values of around zero represent relatively constant TTT measurements over the course of the experiment. The categorical predictions shown in Figs. 4 and 8 are derived by taking samples from the prior distribution over parameters, calculating the corresponding value of $F_{G,u}$ for each sample, and then assigning the result to the category "increase" if $F_{G,u} > 0.05$, "decrease" if $F_{G,u} < -0.05$, and "constant" otherwise. In this way, Figs. 4 and 8 show our a priori expectations of the general trend of TTT development, derived from our prior assumptions. The thresholds of $\pm$ 0.05 were chosen after inspecting the distribution of possible values of $F_{G,u}$ under the priors assumed here, for different G and u. We chose thresholds of ± 0.05 so as to make the prediction "constant" not too likely and not too unlikely over all. Tweaking this threshold may improve model predictions but we refrain from this here, since we presently lack a proper and precise quantitative linking hypothesis for the mapping between posterior odds and properties of mouse trajectories, which could justify any particular choice or inform a data-driven estimate of this parameter.