Cognition 204 (2020) 104372

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Listeners' adaptation to unreliable intonation is speaker-sensitive

Timo B. Roettger[a,*], Kim Rimland[b]

[a] University of Osnabrück, Institute of Cognitive Science, Wachsbleiche 27 (building 50), 49090 Osnabrück, Germany
[b] University of Cologne, IfL-Phonetik, Herbert-Lewin-Str 6, 50931 Köln, Germany

ARTICLE INFO

Keywords:
Prosody
Intonation
Speech adaptation
Mouse tracking
Rational analysis

ABSTRACT

Variable linguistic environments require the ability to quickly update expectations and behavior including speech comprehension. This adaptive capacity is key to understanding how listeners successfully recognize speaker intentions in light of the ubiquitous variability in speech. The present study investigates how listeners' real-time sentence comprehension adapts to speaker-specific prosodic variability. In two forced choice mouse tracking experiments, listeners had to identify a visual referent guided by pre-recorded instructions. When exposed to a speaker that uses unconventional pitch accent placement, listeners discard intonational information for that speaker, but keep using intonation to resolve the referential ambiguity for another speaker that places pitch accents conventionally. These results show for the first time that intonationally guided sentence comprehension adapts in a speaker-sensitive way. The study further provides valuable first insights into the temporal unfolding of this adaptation process. Listeners first attribute unconventional patterns to the context, thus discarding the informational value of intonation for both speakers. After sufficient evidence, however, listeners start attributing unexpected patterns to only the unconventional speaker. Materials, data, and scripts can be retrieved here: http://osf.io/fdpg4

1. Introduction

Human communication is a complex information transmission process that allows us to express our intentions, our emotional states, and our social identity. For example, the English sentence “Bob wrote a book” can be pronounced in many different ways. By modulating acoustic dimensions such as duration, loudness, and pitch of different parts of the utterance, we can express a statement or a question, contrast “Bob” or “book” to already mentioned referents, or express outrage or content (e.g. Ladd, 2008). The transmission of these communicative functions itself, however, is tremendously noisy. The acoustic realizations of communicative units vary quite substantially within and across speakers. This is true for individual segments such as the way we pronounce ‘bear’ vs. ‘pear’, as well as the suprasegmental features of speech such as the way we rise in pitch to highlight something important in a sentence. This “lack of invariance” (Liberman et al., 1967) is arguably a fundamental challenge to speech perception which listeners have to overcome for successful communication (e.g. Weatherholtz and Jaeger, 2016, for a recent review). A large body of work suggests that a successful model of speech perception needs to allow past experiences to influence future input. In line with this idea, listeners have been shown to swiftly update their expectations about a speaker's production behavior in a given communicative context (e.g. Clayards et al., 2008; Dahan et al., 2002; Hay et al., 2006; Idemaru and Holt, 2011; McMurray and Jongman, 2011; Reinisch and Holt, 2014; Ryskin et al., 2019; Trude and Brown-Schmidt, 2012) and use information about a speaker to interpret the acoustic signal (e.g. Creel et al., 2008; Goldinger, 1998; Kraljic and Samuel, 2007; Liu and Jaeger, 2018). Based on previous experiences with similar speakers, listeners can then identify higher level categories such as dialects and adjust future interpretations of the speech signal (Baese-Berk et al., 2013; Bradlow and Bent, 2008). While many models of speech perception share the core assumption that past experiences guide perception of future input, the exact mechanisms and computational assumptions vary across frameworks. We henceforth use the term ‘adaptation’ to refer to the ability to rapidly adjust mappings between linguistic input and cognitive representations through implicit learning.

The ability to adapt to variable environments is an essential capacity of human agents and is central to our understanding of speech communication. Our ability to rapidly update expectations in light of recent input can be accounted for by a variety of cognitive models including episodic and exemplar-based models (e.g. Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2001), as well as rational analysis accounts (e.g. Kleinschmidt and Jaeger, 2015), all of which assume that

* Corresponding author.
E-mail addresses: timo.b.roettger@gmail.com (T.B. Roettger), Kim@Rimland.de (K. Rimland).

https://doi.org/10.1016/j.cognition.2020.104372
Received 18 November 2019; Received in revised form 19 May 2020; Accepted 5 June 2020

experiences with language are stored along with knowledge about the context in which they occur, including information about the interlocutor. The latter family of accounts, rational analysis accounts (Anderson, 1990; Chater and Oaksford, 1999), have been particularly successful in explaining human adaptation behavior in the domain of language processing (e.g. Degen and Tanenhaus, 2015; Frank and Goodman, 2012; Franke, 2009; Kleinschmidt and Jaeger, 2015). These models assume that language users hold probabilistic expectations about the speaker's behavior and continuously draw rational inferences about both speakers' communicative intentions and their mapping between signal and meaning. The proposed inferences are rational because listeners are assumed to integrate prior information about how linguistic variability is generated, probabilistically take uncertainty into account, and draw optimal inferences given the available information. In other words, rational listeners infer the most likely explanation of the linguistic signal they hear.

For rational inferences to be useful, i.e. to successfully attribute variability to contextual factors, there must be systematic structure in the observed variability across environments (e.g., Chodroff and Wilson, 2017; Kleinschmidt et al., 2018; Sumner et al., 2014; Sumner and Samuel, 2009). For example, listeners are expected to update their expectation about how certain vowel categories are pronounced based on sociolinguistic categories such as gender, age, and dialect because these categories systematically vary with respect to vowel pronunciation (e.g. Kleinschmidt et al., 2018). In contrast, listeners are not expected to trace vowel variability as a function of the speaker's t-shirt color, because listeners do not expect a systematic relationship between t-shirt color and vowel formants. Rational listeners, thus, will not condition their inferences just on any covariation, but only on covariation that is expected to be useful in their environment. This naturally raises the question as to the scope of speech adaptation. What aspects of linguistic variation are useful enough to be tracked and eligible for adaptation mechanisms? For example, it remains unclear how listeners deal with competing inferences about possible sources of variability and how this negotiation unfolds as more and more evidence becomes available.

The nature of systematic phonetic variability across segmental patterns of speech (e.g. the difference between /ba/ and /pa/) is well researched and the literature is ripe with research on adaptation to these aspects of speech. However, we cannot draw such a clear empirical picture for prosodic patterns of speech. There is only little work on the adaptation to variability of prosody. This raises the question as to whether and if so how listeners adapt to variability in this domain.

1.1. Variability in intonation

Speakers use prosody, i.e. rhythmic and melodic aspects of speech, to express pragmatic, social, and indexical meanings. For instance, speakers modulate fundamental frequency (perceived as pitch) to signal communicative functions (Bolinger, 1989; Ladd, 2008). We henceforth refer to utterance-wide pitch modulation expressing non-propositional meaning as 'intonation'. Intonation not only allows language users to structure utterances, express their intentions, emotional states and identities, intonation also facilitates efficient speech processing (e.g. Braun et al., 2019; Dahan et al., 2002; Ito and Speer, 2008; Kurumada et al., 2014a; Weber et al., 2006). While playing a central role in human interaction, there is still much unknown about what concrete aspects of the speech signal listeners attend to and how they integrate this information to infer speakers' mental states dependent on the context it occurs in.

A large body of research has attempted to categorize intonational cues in terms of their communicative functions (e.g. Cruttenden, 1997; Cutler et al., 1997; Dahan, 2015; Gussenhoven, 2004; Ladd, 2008; among many others). For example, West Germanic languages have been described as expressing discourse relevant functions by pitch accents, i.e. intonational events co-occurring with lexically stressed syllables (e.g. Ladd, 2008). In English and German, the placement of pitch accents, i.e. which words carry a pitch accent and which do not, and the form of pitch accents, e.g. a falling vs. a rising pitch accent, can signal a referent as discourse-given or as contrastive to a discourse-salient alternative (e.g. Pierrehumbert and Hirschberg, 1990; Baumann, 2006). In (1) below, BOOK, carrying a rising-falling pitch accent, contrasts the sentence object with a set of alternatives (e.g. essay). If the word is unaccented as in (2), it is likely interpreted as a referent that has already been mentioned in the discourse.

1) Bob wrote a BOOK.

→ *It was not an essay that Bob wrote.*

2) BOB wrote a book.

→ *It was not Mary who wrote a book.*

Abstractionist models of intonation attempt to describe systematic relationships between sound and meaning in an informationally efficient way, i.e. they collapse any type of variability across contexts into a compact inventory of categories. Unfortunately, as soon as production data from multiple speakers are considered, such categorization approaches can only account for a limited amount of observations (e.g. Cangemi et al., 2015; Grice et al., 2017). Speakers vary from each other in how they use intonation across dialects and sociolects (e.g. Clopper and Smiljanic, 2011; Holliday, 2019; Warren, 2016) and even within the same language variety (e.g. Grice et al., 2017; Ito, Speer, & Beckman, 2004; Peppé et al., 2000; Turnbull et al., 2017). This variation manifests itself in the way intonational categories are phonetically instantiated, e.g. whether a rising pitch movement reaches its peak a little earlier or later relative to the segmental carriers (e.g. Cole and Shattuck-Hufnagel, 2016). Moreover, speakers vary in their choice of intonational categories to express specific communicative functions (e.g. Grice et al., 2017; Holliday, 2019), e.g. whether they use a falling or a rising pitch movement. Although there are reports that speakers can even differ in the position of intonational events within the utterance (Ito et al., 2004; Roettger, 2017), pitch accent placement, i.e. which words carry a pitch accent, seems to be the most stable aspect of intonational encoding of meaning. In fact, listeners make ample use of pitch accent placement to predict upcoming information (Dahan et al., 2002; Ito and Speer, 2008; Kurumada et al., 2014a; Weber et al., 2006).[1] It is important to emphasize that the large amount of observed variability should not be mistaken as evidence against a systematic mapping between intonation and meaning. Users of intonation languages clearly have systematic knowledge about how to use intonation to express communicative functions. This knowledge is apparent in the probabilistic relationships between meaning and the placement and type of pitch accents, as well as their concrete phonetic realization. For example, variability in categorical pitch accent placement and pitch accent type can be related to systematic differences in the gradual modulation within these categories (Grice et al., 2017; Roessig et al., 2019).

In line with the large amount of variability, perceptual assessments of how listeners map intonation onto speaker intentions indicate rather poor accuracy and substantial flexibility in what constitutes a valid mapping (e.g. Cangemi et al., 2015; Roettger et al., 2019). The fact that there is strong evidence for unaccounted within-speaker variability on the one hand, but mostly stable pitch accent placement across speakers on the other, raises the question as to whether listeners adapt to

[1] Although we consider the mapping of meaning onto pitch accent placement *more* consistent than for example the mapping of meaning onto the concrete form of a pitch accent, it is important to note here that there is still only little empirical evidence to quantify this intuition.

speaker-specific pitch accent placement. It is thus important to carefully investigate how listeners integrate intonational information from speakers who differ in the way they use accent placement.

1.2. Adaptation to unreliable intonation

Recent evidence suggests that listeners can rapidly adapt to recent experiences with varying intonational contours (e.g. Buxó-Lugo and Kurumada, 2019; Kurumada et al., 2014b; Kurumada et al., 2018; Roettger and Franke, 2019). Buxó-Lugo and Kurumada (2019), investigated American English echo questions and statements which can be solely differentiated by the final pitch movement in the utterance. The authors created a continuum between two naturally produced contours and exposed different groups of listeners to different distributions of these contours mapping onto the two discourse functions. After exposure, listeners had to categorize tokens on the continuum as either a question or a statement. The authors showed that listeners can learn to adapt their expectations about the mapping of concrete pitch trajectories onto discourse functions according to recent input (see also Kurumada et al., 2018).

Beyond the adaptation to recent input, the reliability of intonational cues affects listeners' real time integration of them. Otherwise uninformative cues can be learned to be informative if sufficiently reliable in the input. In Roettger and Franke (2019), German listeners first heard a discourse-setting question introducing a referent (see examples 3–4 in Method section) and then heard an answer to this question, either confirming the already mentioned referent or contrasting the mentioned referent with a new one (see examples 5–6 in Method section). Within the microcosm of the experiment, the discourse status of the target referent as given or new could be systematically identified by the presence or absence of an early pitch accent on the verb. Listeners had to choose the correct discourse referent in a two-alternative forced choice task via a mouse click. By analyzing the trajectory of listeners' mouse movements, the authors were able to identify the point in time when listeners started to anticipate the correct discourse referent prior to lexical disambiguation.

Some of these anticipatory patterns only emerged after listeners had encountered a sufficient amount of evidence to learn the association between pitch accent placement and meaning, suggesting a rather swift updating mechanism. Likewise, otherwise highly informative cues can be unlearned if the speaker is sufficiently unreliable (Kurumada et al., 2014b; Roettger and Franke, 2019). When exposed to a speaker that occasionally used unexpected pitch accent placement, listeners in Roettger and Franke's study systematically downgraded the informational value of that accent pattern, directly affecting the time they needed for successful reference resolution.

Similarly, 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 similar discourse functions in the past. Listeners were exposed to a speaker that either uses conventional or unconventional pitch accent placement to mark contrastive focus. After this pre-exposure, listeners' eye movements were monitored during a forced choice task in which listeners heard pre-recorded sentences specifying a referent. One of the intonational constructions contained a contrastive pitch accent that allowed to anticipate the correct referent before lexical disambiguation. The authors showed that those listeners that were pre-exposed to an unconventional speaker, disregarded the early contrastive pitch accent as a cue to the upcoming referent.

These studies provide evidence that experiences with intonational form-function mappings (here accent placement), are tracked along with knowledge about the context they are used in. This interpretation is in line with corpus studies suggesting that linguistic units might be stored alongside their discourse functions and their intonation contours (e.g. Calhoun and Schweitzer, 2012; Schweitzer, 2011).

Studies on adaptation patterns in intonation, however, used one and the same speaker to investigate adaptation, raising the question as to what sources listeners really attributed the unreliability to. There are several plausible scenarios. Given that the mapping between intonation and meaning is variable across speakers, it is plausible to assume that listeners attribute the observed variability to the specific speaker they encountered in the experiment ("speaker-specific adaptation"). This would mean that they do not generalize the observed patterns to other speakers or contexts.

Alternatively, it is equally plausible to assume that listeners generally start with the assumption that speakers use accent placement in a reliable and conventional way. Divergences from these conventions could then be attributed to other factors instead, such as the experimental context ("context-specific adaptation"). In that case, listeners generalize the observed patterns to other speakers within the same context.

These two scenarios are not mutually exclusive. Given that speech systematically varies according to different aspects of its linguistic and non-linguistic environment, it is plausible that observed variability can be attributed to multiple sources at the same time. In other words, listeners might be sensitive to both speaker-specific and context-specific behavior and the attribution of the variability is negotiated between the two sources.

The present study examines whether, and if so, how listeners adapt to speaker-specific intonational variability. When confronted with two distinct speakers that use accent placement differently, do listeners change their online sentence comprehension contingent on their knowledge about individual speakers' behavior or do they change across the board?

2. Method

In two experiments, German participants listened to two different speakers. In the first experiment, both speakers used intonation to mark the discourse status of a referent (given or new) according to the grammatical conventions in German. In the second experiment, one speaker used intonation according to the conventions of German, while the other speaker used intonation in an unexpected way, flipping the conventional mapping between intonation and discourse interpretation. Following previous work (Roettger and Franke, 2019), the stimuli were characterized by intonation contours that allow listeners to anticipate the upcoming referent prior to lexical disambiguation. In a forced choice reference resolution task, listeners had to indicate the correct referent as fast as possible. Tracking the coordinates of their mouse trajectory enabled us to investigate listeners' real time anticipatory use of intonational cues (Roettger and Franke, 2019; Tomlinson et al., 2017).

2.1.1. Participants and procedure

Sixty German subjects participated in the study (24 males, mean age = 25.6 (SD = 4.5)). There were thirty subjects in each experiment. All participants had self-reported normal or corrected-to-normal vision and normal hearing. They were recruited from the Cologne area in Germany and were compensated monetarily. Participants were told about a fantasy creature called 'Wuggy', which picks objects off the ground. There were twelve different objects that the wuggy could pick up (see Fig. 1C).

On each trial, participants heard either a topic question like (3), which introduced a referent as given in the discourse, or a neutral question like (4) introducing no specific discourse content:

3) Hat der Wuggy dann die Geige aufgesammelt?

Did the wuggy then pick up the violin?

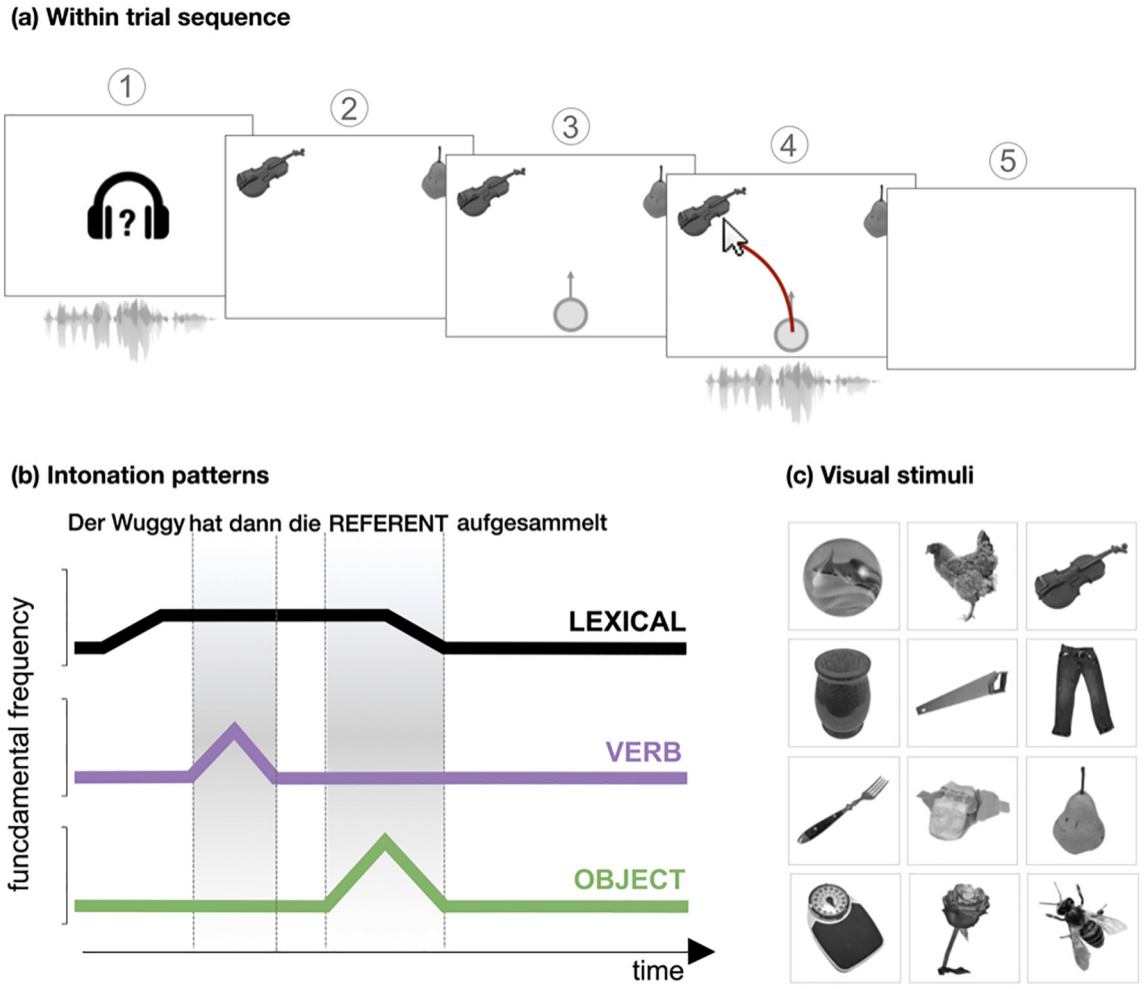

Fig. 1. (a) Schematic depiction of experimental trials. On the question screen (1), participants heard the context-setting question. After a 1000 ms preview of target and competitor referents (2), the initiation button was displayed at the bottom center of the screen (3). Upon clicking the initiation button, listeners started the audio playback of the response sentence, indicating the target referent (4). The trial ended when the cursor was moved into the response box around the target referent, followed by a 1000 ms inter-stimulus interval displaying a blank screen (5). (b) Schematic depiction of the intonation contours characterizing the three different discourse contexts. Grey shading marks relevant landmarks of the intonational cues (presence vs. absence of pitch accent) (c) Visual stimuli (see http://osf.io/drhxy for color versions).

4) Was ist passiert?

What happened?

Following the context screen, participants saw a response screen with two visually presented referents in the upper left and right corner, respectively (left/right placement of target vs. competitor was counterbalanced within participants and items). A click on the start button initiated the audio recording of a statement specifying which object was picked up.

5) Der Wuggy hat dann die Geige aufgesammelt.[2]

The wuggy has then the violin picked-up (literal translation).
The wuggy then picked up the violin.

6) Der Wuggy hat dann die Birne aufgesammelt.

The wuggy has then the pear picked-up (literal translation).
The wuggy then picked up the pear.

Statements were realized with different intonation contours according to intonational conventions in German (e.g. Féry and Kügler, 2008; Grice et al., 2017). After a neutral question (4), a valid answer would be prosodically characterized by a rise in pitch on the subject, followed by high plateau and a fall in pitch preceding the referent. Since no concrete discourse context is given in the question, listeners have to wait until the LEXICAL information becomes acoustically available. After a topic question (3), the utterance in (5) can prosodically emphasize that the proposition in question is true, which is characterized by a high rising accent on the VERB "hat" (engl. *has*). This pitch accent is conventionally interpreted as marking verum focus and thus as confirming the proposition under discussion, in turn indirectly cueing the givenness of the upcoming referent. The answer in (6) negates the topic question (3) and affirmatively mentions a contrastive referent, which is typically characterized by a high rising accent on the sentence OBJECT Birne (engl. *pear*). In this condition, the absence of a pitch accent on the verb can cue the contrastiveness of the upcoming referent (see discussion below). All possible statements (n = 12) came with these three intonation contours (LEXICAL, VERB, and OBJECT), resulting in 36 different target sentences overall (see Fig. 1B for a

[2] The sentences contained the adverb "dann" in order to increase the temporal interval between the verb and the sentence object in the hope to maximize our chances to find a temporal signature of predictive movement behavior.

schematic illustration of the three intonation contours). Each target sentence was repeated twice per speaker, resulting in a total of 144 target trials (12 items * 3 conditions * 2 repetitions * 2 speakers). All trials were split into 12 blocks with 12 trials each. Participants were not aware of the underlying groupings and they solely served the purpose of enabling pseudo-randomization.

Participants were instructed to choose the correct referent as quickly as possible. Prior to the experimental trials, participants familiarized themselves with the paradigm during 12 practice trials. Item pairs and their combination with the experimental conditions were pseudo-randomized for each block. Each referent occurred as a target referent only once per block, i.e. on average every 12 trials. Order of trials within a block and order of blocks were randomized for each participant. There were two groups. In the CONTROL group, both speakers used the conventional intonational cues to indicate discourse-relationships. In the test group, one speaker used intonational cues in a CONVENTIONAL way (counter balanced), while the other used the opposite pattern, which is UNCONVENTIONAL (i.e. VERB for contrastive and OBJECT for given referents).

2.1.2. Materials

Visual stimuli were taken from the BOSS corpus (Brodeur et al., 2010). Recordings of three trained phoneticians served as acoustic stimuli. One male speaker produced the questions. Both a male and a female speaker produced natural answers congruent with the prompting question. To ensure that sentences across the three different intonation contours exhibit comparable durational characteristics, utterances were manipulated and resynthesized using Praat (Boersma and Weenink, 2016) applying the procedure described in appendix A1 (the online repository contains both original and resynthesized stimuli at http://osf.io/drhxy).

2.1.3. Technical set-up

The experiment was created and run in the experiment builder software OpenSesame (Mathôt et al., 2012), using the mousetrap-os plugin (version 2.0.0) (Kieslich and Henninger, 2017) in order to record the streaming x,y-coordinates of the participants' hand movements. The experimental files can be retrieved here: http://osf.io/drhxy. Participants were individually seated in front of a Mac mini 2.5 GHz Intel Core i5. Auditory speech stimuli were presented via a pair of AKG K271 MK II closed-back headphones. Participants controlled the experiment via a Logitech B100 corded USB Mouse. Cursor acceleration was linearized and cursor speed was slowed down (to 1400 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 (Fischer and Hartmann, 2014; Kieslich et al., 2019).

2.1.4. Predictions

Based on previous findings by Roettger and Franke (2019), we expect listeners to anticipate the referent in intonationally informed trials. Listeners anticipate the referent by moving their mouse towards the target image before lexically disambiguating information becomes available. This anticipatory advantage is based on their prior knowledge with German and should already be observable at the beginning of the experiment. When listeners then encounter unconventional form-function mappings in the test group, the temporal advantage at the beginning of the experiment for the CONVENTIONAL speaker might change.

If there is *no adaptation* or a strict *speaker-specific adaptation*, we expect listeners to use accent placement for the CONVENTIONAL speaker in a way similar to how they use accent placement in the CONTROL group. UNCONVENTIONAL trials are solely attributed to the UNCONVENTIONAL speaker and do not affect listeners' performance in CONVENTIONAL trials.

If listeners attribute the unreliable use of intonation *solely to the experimental context*, we expect listeners to rapidly disregard accent placement for the CONVENTIONAL speaker. At the end of the experiment, the temporal advantage in intonationally informed trials should be lost and performance should be comparable to lexical disambiguation.

If adaptation is merely *speaker-sensitive*, listeners remain to some extent uncertain about the source of variability. The inference made about the UNCONVENTIONAL speaker affects the ways in which listeners process the input of the CONVENTIONAL speaker. Thus, the initial temporal advantage in intonationally informed trials should decrease over the course of the experiment. However, listeners modulate their expectation for the speakers to different degrees. Thus, we expect that even at the end of the experiment, there should still be a temporal advantage in CONVENTIONAL trials.

2.1.5. Data analysis

The x, y screen coordinates of the computer mouse were sampled using the mousetrap plugin (Kieslich and Henninger, 2017) implemented in OpenSesame (Mathôt et al., 2012). Trajectories were processed with the R package mousetrap (Kieslich and Henninger, 2017) using R (R Core Team, 2019).

For each trial, we computed the following measurement based on space-normalized trajectories (i.e. trajectories resampled to 101 steps separated by a within-trial constant time interval). We look at the moment in time relative to the unfolding speech signal at which a mouse trajectory starts to migrate uninterruptedly towards the target interpretation. We define the turn-towards-the-target (TTT) as the latest point in time at which the trajectory did not head towards the target horizontally (see Roettger and Franke, 2019). To estimate listeners' adaptation behavior, we used Bayesian parameter estimation based on hierarchical linear regression models. The model fitted turn-towards-target measurements as a function of discourse condition (LEXICAL, VERB, OBJECT), trial group (CONTROL, CONVENTIONAL, UNCONVENTIONAL) as well as their three-way interactions with both a linear and a quadratic term for TRIAL numbers (1–72)[3] using the R package brms (Bürkner, 2016). We included random intercept for both listeners and items as well as by-listener random slopes for the interaction of CONDITION and TRIAL number (see http://osf.io/tyq8v for details on model specifications).

We used weakly informative Gaussian priors centered around zero with $\sigma = 200$ ms for all population-level regression coefficients; truncated Gaussian priors centered at zero for all standard deviations; and LKJ(2) priors for all correlation parameters. These priors are what is referred to as weakly informative or regularizing (Gelman et al., 2008), i.e. our prior assumption is agnostic as to whether the predictors affect the dependent variable, thus making our model conservative with regards to the predictors under investigation.

We operate within the Bayesian inferential framework (rather than within a frequentist framework) for two reasons: First, Bayesian methods allow us to directly answer the primary research question: How plausible is our hypothesis given the data? We can answer this question by quantifying uncertainty about the parameters of interest, which frees us from committing to hard cut-off points for statistical significance (such as the arbitrary 0.05 alpha level). Second, Bayesian inference allows us to flexibly define hierarchical models (also known as mixed effects or multilevel models). Frequentist linear mixed models have become standard in quantitative psychology and they are commonly fit with the R package lme4 (Bates, Mächler, et al., 2015). However, linear mixed effects models that also include complex random effect structures justified by the design (Barr et al., 2013; Bates,

[3] For the test group with two diverging speakers, trial number represents encountered trials with the unconventional speaker in order to track adaptation behavior more accurately. For the control group, trial number simply represents even trial numbers.

Kliegl, et al., 2015; Schielzeth and Forstmeier, 2008) tend not to converge or to give unrealistic estimates of the correlations between random effects (Bates, Kliegl, et al., 2015). In contrast, complex random effects structure can be fit without problems using Bayesian hierarchical models. Compared to converging models with simple random effect structures using lme4, our Bayesian models can be considered more conservative.

For relevant predictor levels and contrasts between predictor levels, we report the posterior probability for turn-towards-the-target values. The posterior distribution is the combination of the prior distribution and the likelihood function derived from the data. Contingent on our model, our data, and our priors, posterior distributions represent our best guess about how the predictors affect the dependent variable. We summarize these distributions by reporting the posterior mean and the 95% credible intervals (henceforth CrIs, calculated as the highest posterior density interval).

3. Results and discussion

Following previous work (Roettger and Franke, 2019), the whole data set of a participant was excluded whenever they (a) exhibited more than 10% errors ($n = 2$), or (b) exhibited initiation times above 350 ms in more than 15% of the trials ($n = 1$). Trials with initiation times greater than 350 ms (1.3%) and incorrect responses (1%) were discarded on a trial-by-trial basis. Initiation times were restricted in order to avoid undesired movement strategies (i.e. wait until the acoustic cue becomes available and move only then).

The following report describes the results using posterior means and 95% CrIs for all relevant predictor levels and contrasts between levels. A 95% CrI demarcates the range of values that comprise 95% of the probability mass of our posterior beliefs in a specific parameter or contrast and can be interpreted as our best guess about where these values lie, informed by the data, the model, and the priors. We consider the evidence for an effect compelling if the value 0 is not included in the 95% CrI. We will focus on the differences between the intonationally informative conditions (VERB, OBJECT) and the lexical disambiguation as probed by the LEXICAL condition. We further compare these differences at the beginning (i.e. at the first point in time listeners encounter any evidence) to the end of the experiment (on the last trial). We refer to a turn-towards-the-target time that is compellingly lower than the time of lexical disambiguation (LEXICAL) as a “predictive advantage”.

Fig. 2 and Table 1 illustrate the experimental results. Looking at the development of turn-towards-the-target measures in Fig. 2a (left panel), the control group serves as a baseline. In the VERB condition (purple), listeners anticipate the referent from the start ($\Delta^{\text{beginning}}$VERB-LEXICAL: −266, 95% CI [−328, −204]). They turn towards the target shortly after hearing the pitch accent on the verb, thus immediately integrating the pitch accent.

In the OBJECT condition (green), listeners also anticipate the referent at the beginning of the experiment, albeit in a much more delayed way ($\Delta^{\text{beginning}}$OBJECT-LEXICAL: −50 [−112,13]),[4] and they become substantially faster over the course of the experiment, leading to a noticeable predictive advantage by the end of the experiment (Δ^{end}OBJECT-LEXICAL: −167 [−241, −98], see Fig. 2b, left panel). In line with Roettger and Franke (2019), we assume that prior to the experiment, the absence of a pitch accent on the verb is not a reliable (enough) cue, but within the microcosm of the experiment, listeners learn to use the absence of a pitch accent on the verb. This development seems non-linear with rapid adjustments over the first half of the experiment followed by a plateau towards the end (confirming observations by Roettger and Franke, 2019).

Looking at listeners' responses to the CONVENTIONAL and the UNCONVENTIONAL speakers in the test group, we observe that the predictive advantage of intonationally informed decisions collapses for the UNCONVENTIONAL speaker towards the end of the experiment (lines overlap heavily).[5] The VERB cue does actually have a predictive *dis*advantage at the beginning of the experiment ($\Delta^{\text{beginning}}$VERB-LEXICAL: 84 [−5174]). This is not surprising because listeners might direct their cursor to the competitor referent based on the expected intonational cue and when hearing the mismatching lexical information, they have to correct their initial response direction. Throughout the experiment, listeners quickly discard the predictive value of intonation. At the end, listeners wait until the referent noun becomes acoustically available (Δ^{end}VERB-LEXICAL: −17 [−114,74]).

Importantly, listeners do not entirely discard the predictive value of intonation for the CONVENTIONAL speaker. At the beginning of the experiment, listeners anticipate the referent using the VERB cue ($\Delta^{\text{beginning}}$VERB-LEXICAL: −280 [−365, −192]), quantitatively comparable to the control group. This predictive advantage becomes weaker over the first half of the experiment, but, crucially, stabilizes towards the end. At the end of the experiment, there is still compelling evidence that listeners anticipate the referent based on the VERB cue (Δ^{end}VERB-LEXICAL: −103 [−196, −7]). Listeners turn towards the target much later than at the beginning of the experiment - an indication of the adjusted predictive value of the cue (Roettger and Franke, 2019). Thus, listeners still use the pitch accent information to predict the upcoming referent to some extent. Note that this preservation of a predictive advantage cannot be observed for the OBJECT cue (Δ^{end}OBJECT-LEXICAL: 10 [−86,105]) which can be considered a weak cue (see above).

4. General discussion

Speech communication is characterized by tremendous variability across and within speakers. To accommodate such variable linguistic environments, listeners can adapt to different contextual factors (e.g. Goldinger, 1998; Kleinschmidt and Jaeger, 2015; Pierrehumbert, 2001), including idiosyncratic speech patterns of different speakers (e.g. Kraljic and Samuel, 2007; Trude and Brown-Schmidt, 2012). The present study contributes to our understanding of listeners' adaptation to intonation, which is characterized by variable phonetic instantiation, but, at the same time, exhibits strong systematicity in terms of the discrete placement of intonational events.

In two experiments, German participants listened to two speakers, differing in whether they use intonation conventionally or not. In a forced-choice reference resolution task using the mouse tracking paradigm, listeners had to indicate the correct referent by moving their mouse to the image that depicts the target referent. We measured the time at which listeners start moving their mouse horizontally towards the target image. If this direction change happened before directional changes characteristic of lexical disambiguation, we considered this as evidence for anticipation of the target interpretation using intonational information.

In line with previous work on pitch accent processing (e.g. Dahan et al., 2002; Ito and Speer, 2008; Kurumada et al., 2014a; Watson et al., 2008; Weber et al., 2006), our results indicate that listeners make use of intonation to anticipate reference resolution when encountering only reliable speakers. When encountering one reliable and one unreliable speaker, listeners adapt their anticipatory behavior. Listeners discard intonation as predictively valuable information for a speaker that uses

[4] Given that zero is included in the 95% CrI, we should consider these results as only weak evidence.

[5] An alternative interpretation is that an initial disadvantage of the verb cue collapses. In other words, listeners are *learning* to re-interpret the cue as having the *opposite* of its conventional meaning. To tease apart these interpretations, one would have to increase the trial number to see what happens to listeners' interpretation strategy when further evidence accumulates.

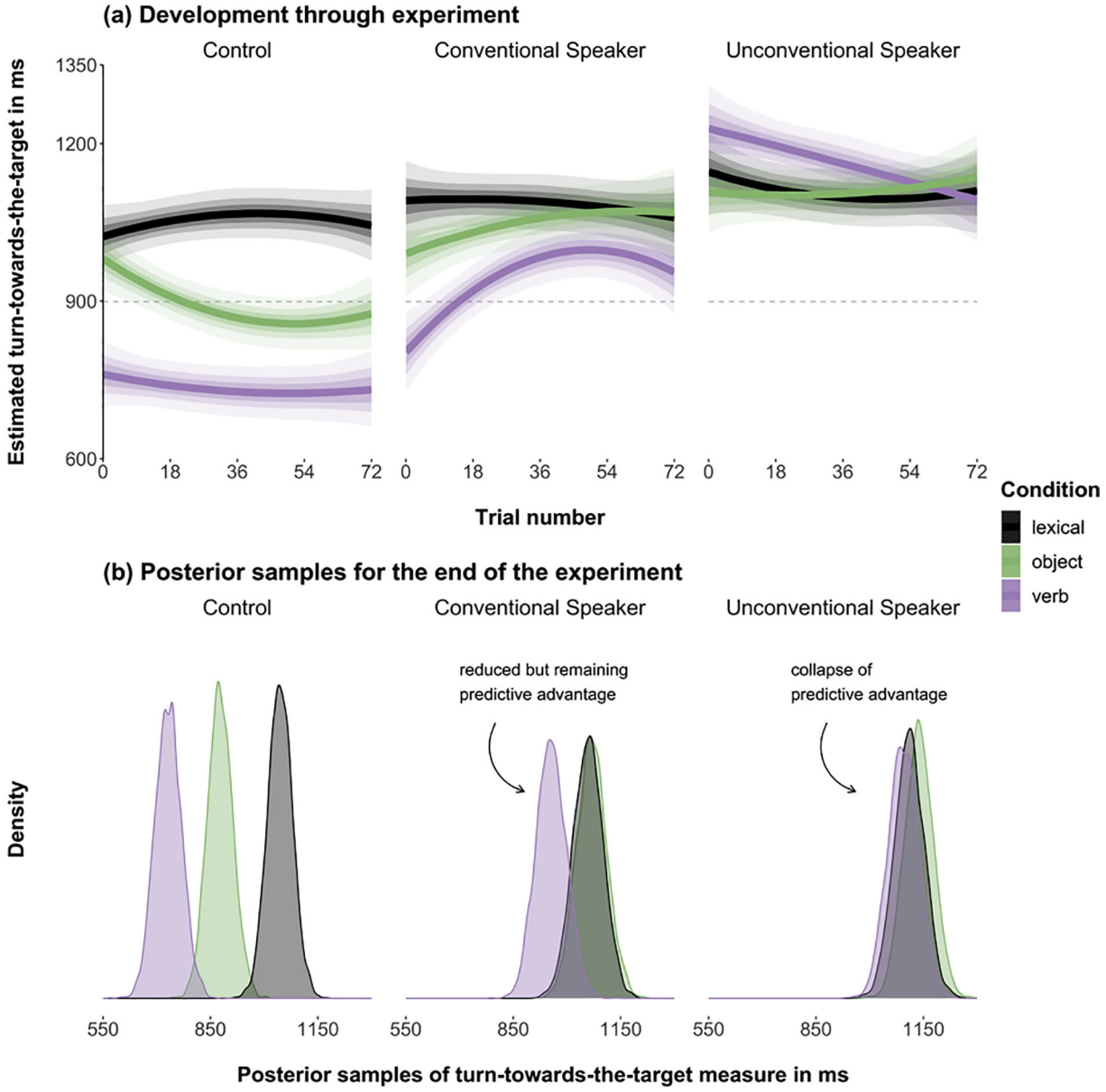

Fig. 2. Model estimates for experimental results. (a) Posterior means and 50%/75%/95% credible intervals (different shading) of turn-towards-the-target measures across conditions, trial number and groups. The dashed line indicates the averaged acoustic onset of the target referent; (b) Posterior samples for the end of the experiment (last trial) across conditions and groups.

Table 1
Experimental results. Posterior means (95% credible intervals) of turn-towards-the-target measures in ms for the differences between LEXICAL disambiguation and intonationally informed decision (VERB and OBJECT, respectively) for groups at the beginning (1st trial) and end of the experiment (last trial). Negative numbers indicate a predictive advantage, i.e. a turn-towards-the-target prior to lexical disambiguation. Grey cells indicate posterior estimates for which the 95% CrIs do not include 0.

Beginning/end of experiment	Group	Condition	Posterior mean [95% CrI]
Beginning	Control	Δobject-lexical	-50 [-112, 13]
		Δverb-lexical	-266 [-328,-204]
	Conventional	Δobject-lexical	-100 [-183,-8]
		Δverb-lexical	-280 [-365,-192]
	Unconventional	Δobject-lexical	-35 [-123,52]
		Δverb-lexical	84 [-5,174]
End	Control	Δobject-lexical	-167 [-241,-98]
		Δverb-lexical	-312 [-381,-242]
	Conventional	Δobject-lexical	10 [-86,105]
		Δverb-lexical	-103 [-196,-7]
	Unconventional	Δobject-lexical	27 [-62,122]
		Δverb-lexical	-17 [-114,74]

intonation unreliably (see also Kurumada et al., 2014b). This process unfolds over the course of the experiment as they encounter more and more evidence. The data suggest that this down weighing process is linear.

For the speaker who uses intonation conventionally, there is a non-linear adjustment of listeners' real-time comprehension strategy. After a few observations, the initial predictive advantage becomes smaller (i.e. turns towards the target become delayed). The decreasing trend continues over the first half of the experiment. After sufficient evidence has been accumulated, however, the predictive advantage 'stabilizes', i.e. stops decreasing. At that point, listeners appear to have learned to attribute the unreliable intonation pattern to one of the speakers (instead of e.g. the experimental context). At that point they keep using intonation in an anticipatory way for the conventional speaker but not for the unconventional speaker. This is to our knowledge the first study providing evidence that listeners integrate their knowledge about speaker identity into the way they process prosody in real-time, adding to the growing literature on speaker-sensitive linguistic processing (e.g. Arnold et al., 2007; Grodner and Sedivy, 2011; Schuster and Degen, 2020; Yildirim et al., 2016).

These findings are in line with rational models of linguistic inferences (e.g. Frank and Goodman, 2012; Franke, 2009; Kleinschmidt and Jaeger, 2015; Roettger and Franke, 2019). This family of models proposes that listeners continuously draw rational inferences about the context, including the speakers' communicative intentions and their mapping between signal and meaning. While the acoustic manifestation of intonational categories is very variable across speakers, the placement of linguistically relevant intonational events to express discourse functions (e.g. Bob WROTE a book vs. Bob wrote a BOOK) is rather stable across the speech community (e.g. Pierrehumbert and Hirschberg, 1990). It is reasonable to assume that listeners expect speakers to produce somewhat systematic pitch accent placement when expressing discourse relationships. When encountering an unexpected scenario in which one speaker follows grammatical conventions and another does not, it is not rational (given the evidence and their prior expectations) to immediately attribute unreliable behavior to speaker identity.

Listeners, if they are indeed behaving rationally, should remain to some extent uncertain about the source of the variability observed in the input and should attribute at least some of it to a different source than speaker identity. This is what we observe. Listeners start parcelling out these sources of unreliability only after encountering enough evidence.

Much remains unclear about how listeners attribute unreliable information to possible sources at the beginning of the experiment. Listeners could simply average over recent experiences. In the test group, the intonation in *all* utterances of the unconventional speaker is unconventional. Thus, across all utterances they hear, the mapping of accent placement and discourse meaning is uninformative (any given accent pattern maps onto a specific interpretation with a chance of 50%). Alternatively, listeners could attribute the unreliable tokens to the experimental context or a sub-category of speakers that is derived in an ad hoc fashion. The experimental design does not allow us to tease these (speculative) alternatives apart. The present study additionally contributes important new insights on how this inferential process unfolds as evidence accumulates. After an initial lumping of speakers, indicative of a speaker-overarching adaptation strategy, listeners start to attribute unreliable patterns in a speaker-sensitive way.

Our study further contributes to work on causal inferences during real-time sentence comprehension and extends it to the domain of prosodic processing. Previous work has investigated causal inference in reference resolution. Using the eye-tracking paradigm, Grodner and Sedivy (2011) presented evidence that otherwise observable pragmatic inferences do not occur when participants were told that the speaker is impaired before the experiment. Arnold et al. (2007) presented evidence that listeners fixate unfamiliar objects more often when the speaker produces disfluent speech, suggesting that listeners took the disfluencies as an indication of lexical access difficulties. When listeners were informed that the speaker had difficulties naming familiar objects, the original effect was substantially reduced. These findings suggest that online reference resolution can be informed by top-down inferences about the speaker and their cognitive abilities. In both studies, listeners were explicitly informed about plausible causes. In Liu and Jaeger (2018)'s study on adaptation in speech perception, listeners were not explicitly informed about plausible causes of a speaker's variable speech behavior. Nevertheless, listeners inferred possible disruptions from visual information about the speaker. Our study possibly contributes to this literature by showing that, in absence of any additional information about the speakers, listeners parcel out different sources of variation solely based on observed evidence.

In conclusion, computational models of intonational processing (e.g. Roettger and Franke, 2019) need to consider different levels of representations and allow for simultaneous adaptation to different contextual factors including speaker identity. Tracking contextual information to optimize the information integration of intonation is important for successful communication because intonational categories are characterized by contextual and speaker-specific variation (e.g. Cole and Shattuck-Hufnagel, 2016; Grice et al., 2017). Moreover, listeners must continuously infer plausible causes of variation rooted in non-linguistic information, e.g. the environment in which they encounter linguistic input. While we tested accent placement, a strikingly systematic aspect of intonation, listeners seem to be able to attribute observed variability to speaker identity after sufficient exposure. This is an important observation. Expectations about domains of linguistic behavior that are not directly observed to systematically vary across speakers can be updated in light of new evidence. This raises the question as to whether listeners have strong prior expectations about linguistic variability in general, enabling them to swiftly update their beliefs in light of any kind of variability, or whether listeners' priors are more bounded to specific linguistic domains. These questions strike us as fruitful avenues to explore in future research.

CRediT authorship contribution statement

Timo B. Roettger: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Kim Rimland:** Data curation, Writing - review & editing.

Acknowledgements

Timo Roettger's work was supported by the "Zukunftskonzept" of the University of Cologne as part of the Excellence Initiative. We would like to thank Chigusa Kurumada, two anonymous reviewers, and the editor for their insightful comments and suggestions. All remaining errors are our own.

Appendix A

A.1. Acoustic manipulation

First, we segmented all stimuli. Segmental boundaries were identified using a waveform and a wide-band spectrogram at (relatively unambiguous) major discontinuities in the acoustic signal. We chose intervals to be as unambiguous as possible, resulting in intervals of different sizes. Intervals were sometimes smaller than a phone (e.g. closure duration and VOT of a stop separately), a phone (e.g. /g/,/a/ ,/b/ of "Gabel" 'fork') or intervals larger than phones (e.g. /dɐ/, /vʊ/, /gi:ha/ of "der Wuggy ha(t)").

Second, we measured the duration of each interval for all sentences and calculate the mean duration of each interval across conditions (e.g.

we take /dɐ/ of the sentence “Der Wuggy hat dann die Birne aufgesammelt” for all three intonation contours and calculate its mean duration).

Third, we changed the duration of all intervals to these calculated mean values. This manipulation resulted in stimuli that have comparable temporal characteristics across prosodic conditions. This procedure is preferred over alternative resynthesis protocols: In Roettger and Franke (2019), different intonation contours were resynthesized from a single base stimulus. This protocol ensured temporal uniformity across intonational conditions but came at the cost of somewhat unnaturally sounding stimuli. The present synthesis procedure results in naturally sounding stimuli (as informally judged by both naïve and trained native German speakers). There remain very small differences between single items across conditions because the speech signal cannot be objectively segmented. However, the overall durational differences across conditions are neglectable and likely below perceptual thresholds. The online repository contains both original and resynthesized stimuli at http://osf.io/drhxy.

References

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

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(5), 914–930 10/cq3fnk.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America, 133*(3), EL174–EL180 10/f4skwv.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015b). *Parsimonious mixed models.* Eprint ArXiv:1506.04967.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015a). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Baumann, S. (2006). *The intonation of givenness: Evidence from German. Vol. 508.* Walter de Gruyter.

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer. [Computer program]. Version 6.0.17.

Bolinger, D. L. M. (1989). *Intonation and its uses: Melody in grammar and discourse.* Stanford University Press.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707–729.

Braun, B., Asano, Y., & Dehé, N. (2019). When (not) to look for contrastive alternatives: The role of pitch accent type and additive particles. *Language and Speech, 62*(4), 751–778 10/ggkhdd.

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), Article e10773.

Bürkner, P.-C. (2016). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Buxó-Lugo, A., & Kurumada, C. (2019). *Encoding and decoding of meaning through structured variability in intonational speech prosody.* https://doi.org/10.31234/osf.io/9y7xj.

Calhoun, S., & Schweitzer, A. (2012). Can intonation contours be lexicalised? Implications for discourse meanings. In G. Elordieta, & P. Prieto (Vol. Eds.), *Prosody and meaning. Vol. 25. Prosody and meaning* (pp. 271–328). de Gruyter.

Cangemi, F., Krueger, M., & Grice, M. (2015). Listener-specific perception of speaker-specific production in intonation. In S. Fuchs, D. Pape, C. Petrone, & P. Perrier (Eds.). *Individual differences in speech production and perception* (pp. 123–145). Peter Lang Publishing Group.

Chater, N., & Oaksford, M. (1999). Ten year of the rational analysis of cognition. *Trends in Cognitive Sciences, 3*(2), 57–65.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics, 61*, 30–47 10/ggktm8.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804–809.

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, 7*(1).

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106*(2), 633–664 10/bh7vfw.

Cruttenden, A. (1997). *Intonation.* 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.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science, 39*(4), 667–710.

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. (2009). *Signal to act: Game theory in pragmatics.* Amsterdam: Institute for Logic, Language and Computation.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., & others. (2008). A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2(4), 1360–1383.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251.

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. (2004). *The phonology of tone and intonation.* Cambridge University Press.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics, 34*(4), 458–484. https://doi.org/10.1016/j.wocn.2005.10.001.

Holliday, N. R. (2019). Variation in question intonation in the Corpus of Regional African American language. *American Speech, 94*(1), 110–130. https://doi.org/10.1215/00031283-7308038.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance, 37*(6), 1939–1956. https://doi.org/10.1037/a0025641.

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. *Proceedings of Speech Prosody, International Conference.* Nara, Japan: ISCA.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. W. Mullennix (Eds.). *Talker variability in speech processing* (pp. 145–165). Academic Press.

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

Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: A practical guide to implementation and analysis. In M. Schulte-Mecklenbeck, A. Kuehberger, & J. G. Johnson (Eds.). *A handbook of process tracing methods: 2nd edition*(2nd ed.). Routledge.

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 10/f678h8.

Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science, 10*(4), 818–834. https://doi.org/10.1111/tops.12331.

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. *Proceedings of the Cognitive Science Society,* 36.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2018). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin & Review, 25*(3), 1153–1160.

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

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431–461 10/dvfs2x.

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition, 174*, 55–70 10/gc9p3p.

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.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*(2), 219–246.

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

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.). *Intentions in communication* (pp. 271–311). MIT Press.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. L. Bybee, & P. Hopper (Eds.). *Frequency and the emergence of linguistic structure* (pp. 137–158). John Benjamins Publishing.

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation

for Statistical Computinghttps://www.R-project.org/.
Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance, 40*(2), 539–555 10/f5zxcz.
Roessig, S., Mücke, D., & Grice, M. (2019). The dynamics of intonation: Categorical and continuous variation in an attractor-based model. *PLoS One, 14*(5), Article e0216859. https://doi.org/10.1371/journal.pone.0216859.
Roettger, T. B. (2017). *Tonal placement in Tashlhiyt: How an intonation system accommodates to adverse phonological environments. Vol. 3*. Language Science Press.
Roettger, T. B., & Franke, M. (2019). Evidential strength of Intonational cues and rational adaptation to (un-)reliable intonation. *Cognitive Science, 43*(7), Article e12745. https://doi.org/10.1111/cogs.12745.
Roettger, T. B., Mahrt, T., & Cole, J. (2019). Mapping prosody onto meaning – the case of information structure in American English. Language. *Cognition and Neuroscience, 34*(7), 841–860.
Ryskin, R., Ng, S., Mimnaugh, K., Brown-Schmidt, S., & Federmeier, K. D. (2019). Talker-specific predictions during language processing. *Language, Cognition and Neuroscience,* 1–16 10/ggkg7f.
Schielzeth, H., & Forstmeier, W. (2008). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology, 20*(2), 416–420.
Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition, 203*, 104285.
Schweitzer, K. (2011). *Frequency effects on pitch accents: Towards an exemplar-theoretic approach to intonation.* PhD ThesisUniversität Stuttgarthttp://elib.uni-stuttgart.de/handle/11682/3011.
Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology,* 4. https://doi.org/10.3389/fpsyg.2013.01015.
Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language, 60*(4), 487–501 10/br8f4z.
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 & Cognitive Processes, 27*(7–8), 979–1001.
Turnbull, R., Royer, A. J., Ito, K., & Speer, S. R. (2017). Prominence perception is dependent on phonology, semantics, and awareness of discourse. *Language, Cognition and Neuroscience, 32*(8), 1017–1033.
Warren, P. (2016). *Uptalk: The phenomenon of rising intonation.* Cambridge University Press.
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.
Weatherholtz, K., & Jaeger, T. F. (2016). *Speech perception and generalization across talkers and accents.* Oxford Research Encyclopedia of Linguistics 10/ggkh7g.
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.
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.