Information integration and adaptation during intonation-based intention recognition

Timo B. Roettger[1,2] & Michael Franke[3]

[1] University of Cologne

[2] Northwestern University

[3] University of Tübingen

Author Note

Correspondence concerning this article should be addressed to Timo B. Roettger, Herbert-Lewin-Str. 6, D-50931 Cologne. E-mail: timo.b.roettger@gmail.com

Abstract

Intonation plays an integral role in comprehending spoken language. It is also remarkably variable, often exhibiting only probabilistic mappings between form and function. Despite this apparent uncertainty, listeners rapidly integrate intonational information to predictively map a given pitch accent onto respective speaker intentions. We use manual response dynamics (mouse-tracking) to investigate two questions: (i) whether listeners draw predictive inferences from the presence and absence of an intonational marking and (ii) how listeners adapt their online interpretation of intonational cues when these are reliable or stochastically unreliable. Our results are compatible with the assumption that comprehenders 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.

Keywords: mouse-tracking, intonation, prosody, speech adaptation, rational predictive processing

Word count: ca. 4800 in main text (w/o references & captions), based on texcount tool for *.tex file

Information integration and adaptation during intonation-based intention recognition

1. Introduction

Intonation plays an integral role in comprehending spoken language. It encodes social functions, expresses speaker involvement, emotions, and attitude, and it plays a crucial role in linguistic organization (Ladd, 2008). In languages such as English and German, for instance, the position and form of a pitch accent can signal a referent as discourse-new or contrastive (e.g., Büring, 2009; Féry & Kügler, 2008; Pierrehumbert & Hirschberg, 1990). Traditional descriptions assume a one-to-one mapping of intonational form and functional interpretation (e.g., Pierrehumbert & Hirschberg, 1990). More recent work identifies intonational form-function mappings as highly variable and probabilistic (e.g., Grice, Ritter, Niemann, & Roettger, 2017; Roettger, 2017 for recent discussions). Despite the variability and stochasticity, comprehenders can rapidly integrate intonational cues during online processing to anticipate a likely speaker-intended referent even before disambiguating lexical material is heard (e.g. *inter alia*, Dahan, Tanenhaus, & Chambers, 2002; Ito & Speer, 2008; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014a,Roettger and Stoeber (2017); Watson, Tanenhaus, & Gunlogson, 2008; Weber, Braun, & Crocker, 2006).

An interesting open issue concerns the differential strength of intonational cues. Most studies on intonational processing have focused on how the presence of an intonational cue can be used to anticipate the speaker's intended meaning. Less attention has been payed to whether the *absence* of an intonational cue can be exploited for predictive processing in a similar way.[1] Dahan et al. (2002) found that listeners

[1] "Is there any point to which you would wish to draw my attention?"

"To the curious incident of the dog in the night-time."

"The dog did nothing in the night-time."

"That was the curious incident," remarked Sherlock Holmes.

(From "Silver Blaze", in "The Memoirs of Sherlock Holmes" by Sir Arthur Conan Doyle)

interpreted deaccented nouns anaphorically (see also Weber et al., 2006). Carbary et al. (2015) showed that anticipatory deaccenting patterns contributed to listeners' referential expectations in a gating task. In contrast, Kurumada, Brown, Bibyk, Pontillo, and Tanenhaus (2014b) did not find any evidence for predictive processing based on the absence of pitch accent information; the authors hypothesize that this may in part be due to the specific experimental design in which absence of the relevant cue was compatible with multiple different interpretations.

Perceptual matters aside, a rational predictive interpreter should not make any categorical difference between absence and presence of cues (Hsu, Horng, Griffiths, & Chater, 2017). If we assume a Bayesian pragmatic interpreter (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016), what matters for rational predictive interpretation are rather 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 comprehender'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 produces u for r_i) and the prior odds (how likely a speaker refers to r_i):

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

All else equal, if utterance u with its specific intonational 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 (Edwin Thompson Jaynes, 2003; Jeffrey, 2002). The amount of observational evidence in favor of an interpretation, i.e., the strength of an intonational cue, would depend on the ratio of production likelihoods, not on a categorical distinction between presence and absence of a cue.

A direct experimental measure of comprehenders' dynamically evolving posterior

odds between two candidate interpretations can be obtained from mouse-movements in a forced-choice decision task. Roettger and Stoeber (2017) have recently shown that listeners integrate intonational information early on and move their mouse towards a likely target referent before they have processed disambiguating lexical information. This is in line with numerous experiments demonstrating that the continuous uptake of sensory input and dynamic competition between simultaneously active representations is reflected in subjects' hand or finger movements (e.g., *inter alia* Dotan, Meyniel, & Dehaene, 2018; Freeman & Ambady, 2010; Magnuson, 2005; Spivey, Grosjean, & Knoblich, 2005) and falls in line with recent papers using mouse tracking to investigate the processing of pragmatic inferences (Tomlinson, Bailey, & Bott, 2013; Tomlinson Jr, Gotzner, & Bott, 2017).

Numerous recent studies document how comprehenders may accommodate their online processing strategies to variable linguistic input, such as in phonology (e.g., Kleinschmidt & Jaeger, 2015), syntax (Fine & Florian Jaeger, 2013; Jaeger & Snider, 2013; Norris, McQueen, & Cutler, 2003) or semantics and pragmatics (Grodner & Sedivy, 2011; Yildirim, Degen, Tanenhaus, & Jaeger, 2016). Looking at comprehending intonation, Kurumada et al. (2014b) investigated listeners' online interpretation of intonational cues after a pre-exposure phase in which speakers used 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. It is *a priori* conceivable that comprehenders learn to exploit reliable cues during the course of the experiment as a form of rational task-adaptation. Conversely, comprehenders might also start with good hopes to expect reliable cues (based on conventional stochastic regularities in speech production), but that they unlearn rapid cue exploitation when certain cues prove

unreliable over time. This issue is important because it address the extent to which we should believe that rational rapid cue exploitation is "just" a task-induced effect or rather a genuine propensity and pre-inclination of language users. The study presented here tries to shed light on this issue using a between-subjects manipulation of the frequency of unreliable input within the experimental trials themselves (see Dennison and Schafer (2010) for a similar, but arguably less subtle manipulation).

The study presented here aims to address the question of how listeners adapt their online interpretation of potential intonational cues dynamically during exposition to either entirely reliable or occasionally unreliable form-function mappings. We use manual response dynamics as a window into comprehenders' posterior odds in favor of one interpretation over another after hearing a partial utterance with a succinct intonational pattern. Our design further allows a direct comparison of the evidential strength of "absent" and "present" cues. Section 2 introduces the experiment. Section 3 describes the results. Section 4 discusses the results and explores a formal model of rational incremental interpretation to explain key qualitative patterns observed in the data.

2. Methods

The following experiment was preregistered on the 4th of July 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/.

2.1 Participants and procedure

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

Subjects were told about a fantasy creature called "wuggy", which carries things around. There were twelve different objects that the wuggy could pick up (bee, chicken,

diaper, fork, marble, pants, pear, rose, saw, scale, vase, violin).

Each trial exposed subjects first to a context screen, which was shown for 2500 ms and provided a specific discourse context. Concretely, participants heard either a *topic question* like (1), which introduced a referent as given in the discourse, or the *neutral question* (2):

(1) Hat der Wuggy dann die Geige aufgesammelt?
Did the wuggy pick up the violin then?

(2) Was ist passiert?
What happened?

Following the context screen, participants saw a response screen with two visually presented response alternatives, each depicting one object in the upper left and right corner, respectively (left/right placement of target vs. competitor response alternatives was counterbalanced within participants and items). After 1000 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, e.g. (3) or (4).

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

Participants were instructed to move their mouse immediately upwards after clicking the initiation button (see Spivey et al., 2005) 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 (straight upward) before

the onset of relevant acoustic information, which enables distinguishing properties in the acoustic signal to influence the continuous motor output during its movement. After each response selection, the screen was left blank for a 1000 ms inter-stimulus interval.

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 1400 sensitivity) using the CursorSense© application (version 1.32). Slowing down the cursor ensured that motor behaviour was recorded as the acoustic signal unfolded resulting in a smooth trajectory from start to target.

Prior to the experimental trials, participants familiarized themselves with the paradigm during 16 practice trials.

Statements were acoustically manipulated to exhibit three different intonation contours (see Figure 1). Depending on the preceding context question (1) or (2), statements in (3) and (4) are prototypically realized with different intonation contours (*inter alia*, Féry & Kügler, 2008; Grice et al., 2017). After a neutral question (2), both subject and object are discourse-new which can be prosodically encoded by specific pitch accents on both constituents (often referred to as *broad focus*). A common contour in these cases is a rising accent on the subject, followed by a high stretch of f_0 and a high or falling accent on the object. After a polar topic question (1), the utterance in (3), which affirmatively picks up the *given referent*, can prosodically emphasize that the proposition in question is true, for example, by *verum focus*, which manifests itself here in the form of a high rising accent on the auxiliary ("hat" "has"). Finally and in contrast to the latter, the answer in (4) negates the topic question (1). It affirmatively mentions a *contrastive referent*, which is typically realized by *contrastive focus*, an intonation contour with a high rising accent on "Birne" "pear"). All possible statements (n = 12) came with these three intonation contours (broad, verum, and contrast), resulting in 36 different target sentences overall.

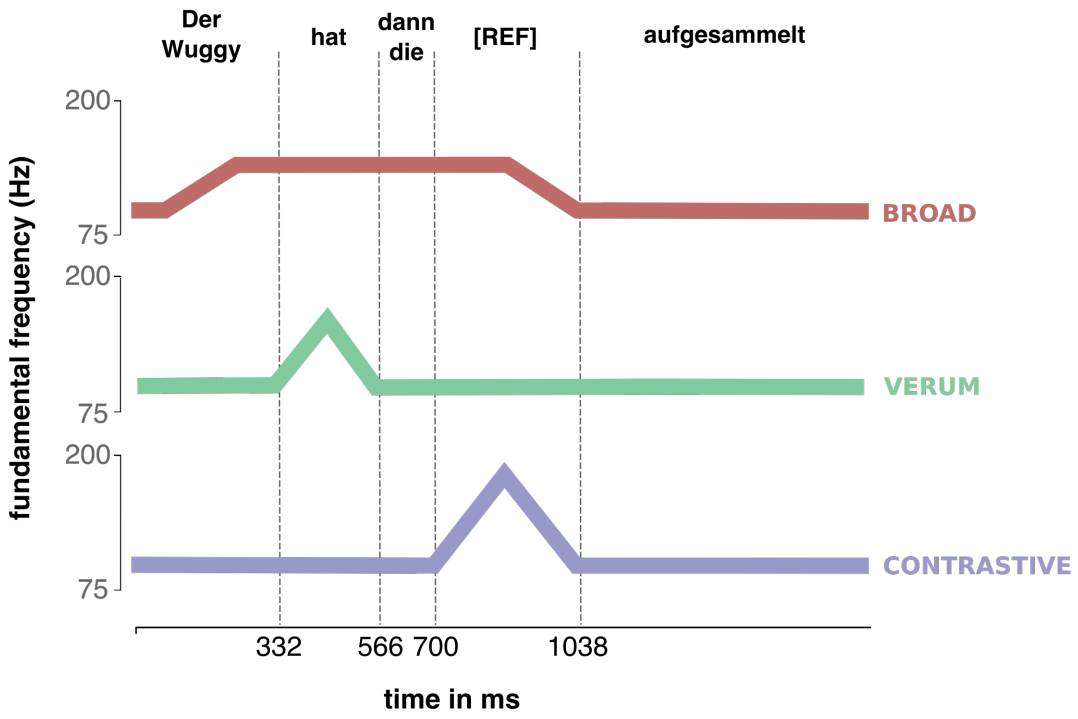

Figure 1. Schematic f_0 contours and average temporal landmarks for the resynthesis of broad, verum and contrastive focus. See supplementary file II for additional details about the resynthesis procedure.

There were two experimental groups. The reliable speaker (RS) group was only exposed to natural 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 phonological form (pitch accent position) and function (the respective discourse status of the referent). In contrast, the unreliable speaker (US) group was sometimes exposed to mismatching intonation. A mismatch occurs when, in the context of a topic question like (1), the speaker uses a statement like (3) realized with a pitch accent on the object as if to indicate a contrastive referent; or a statement like (4) realized with a pitch accent on the verb as if to indicate a given referent. Occasional mismatch leads to a scenario in which listeners cannot fully rely on the speaker's form-function mappings. Subjects were exposed to twelve blocks à eight stimuli. In the

US group, each block contained two contrastive focus statements, two verum focus statements, and four broad focus statements, resulting in 96 trials. Each block in the US group was the same except that there were only two broad focus statements and, additionally, two unreliable mappings (one with mismatching contrastive focus and one with mismatching verum focus). In sum, the US group received additional unreliable trials but less control trials (broad focus) than the RS group.

Each participant was randomly assigned to one reliability group. Item pairs and their combination with focus condition were pseudorandomized for each block. Order of trials within a block and order of blocks were randomized for each participant.

2.2 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 context 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 contexts exhibit the same temporal characteristics for each sentence (i.e. the lexical information become available at the same time across focus conditions), sentences were manipulated and resynthesized using Praat (Boersma & Weenink, 2016). The resulting stimuli differed only in the pitch contour and accompanied intensity envelope. The preregistration report (https://osf.io/dnbuk/) and the supplementary file II contain additional information.

2.3 Data analysis

The x, y 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 were a total of 96 target trials for the RS group. For the US speaker condition, we only analysed the 72 target trials with reliable mappings between discourse context and intonation.

For each trial, we computed two measurements based on time- and space-normalized trajectories. First, overall reaction times (RT), from the initiation click to the target response, serve as a latency baseline. Second, to link manual response dynamics to listener's dynamically unfolding posterior beliefs about likely interpretations 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.[2,3]

3. Results

The whole data set of a participant was excluded whenever he/she (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 subject.

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%). The remaining data went into the statistical analyses.

[2]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 included analysis scripts.

[3]We also measured and analyzed two spatial parameters; see supplementary file I for details.

3.1 Descriptive assesment of trajectories

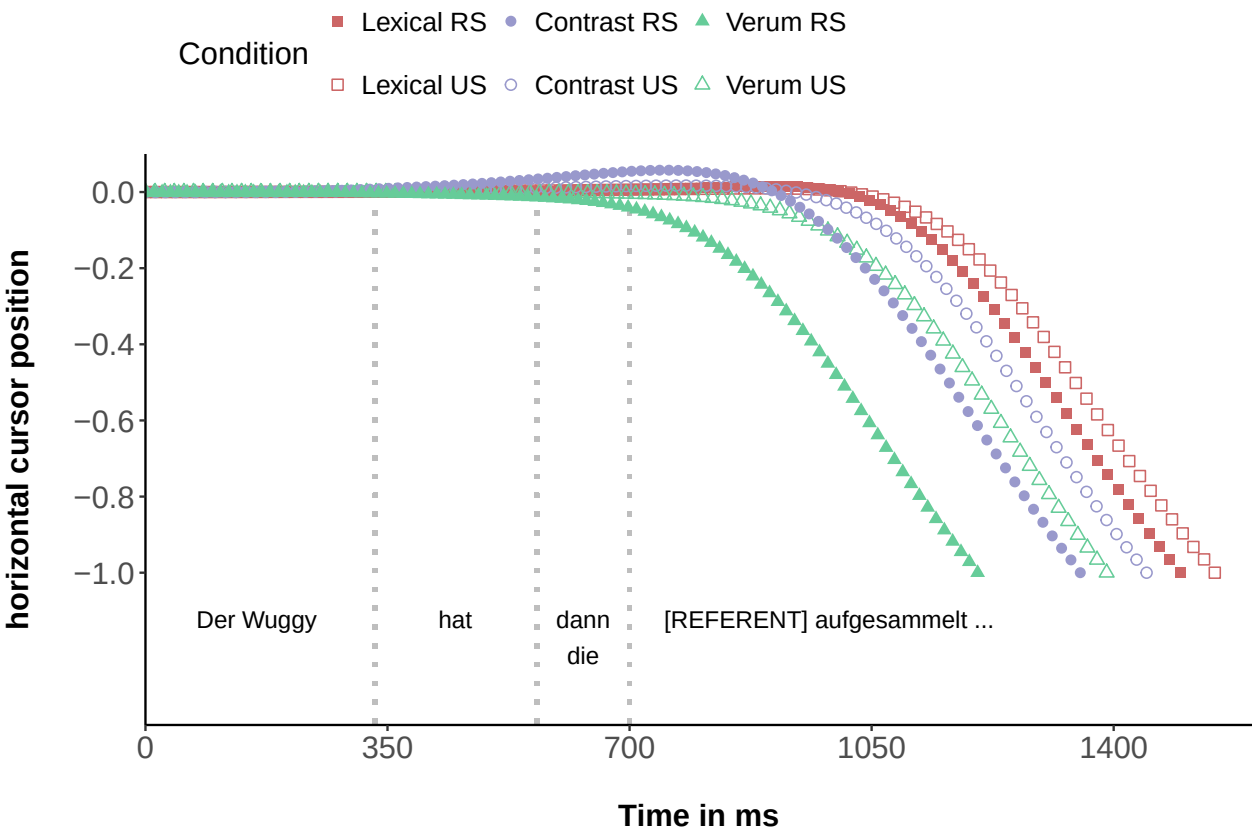

Figure 2. Horizontal cursor position of time- and space-normalized averaged trajectories for the reliable-speaker group (filled symbols) and the unreliable-speaker group (empty symbols).

Figure 2 displays the horizontal cursor position over time as a function of focus and listener groups. Looking at the time course of the decision process, there are clear temporal differences between conditions. In the reliable speaker group (filled symbols), there are strong differences between all three conditions, with the verum focus showing the earliest horizontal turn to the target (y = -1) followed by contrast and broad, indicating the least amount of response competition, and the overall smallest response time.

This temporal pattern is very similar to the unreliable-speaker group (empty symbols), albeit differences in the latter are smaller. Nevertheless, descriptively, verum

focus turns to the target earliest, followed by contrastive focus, and broad focus turns to the target the latest.

3.2 Inferential assessment

We fitted Bayesian hierarchical linear models to reaction times and turn-towards-target measurements as a function of FOCUS, GROUP and BLOCK and their interaction, using the Stan modelling language (Carpenter et al., 2016) and the package *brms* (Buerkner, 2016). The models included maximal random-effect structures, allowing the predictors and their interactions to vary by subjects (FOCUS - BLOCK) and experimental items (FOCUS - BLOCK- GROUP). We used weakly informative Gaussian priors centered around zero with $\sigma = 100$ for all population-level regression coefficients (e.g., Gelman, 2006), as well as standard priors of the *brms* package for all other parameters. Four sampling chains with 4000 iterations each were run for each model, with a warm-up period of 2000 iterations. We report, for each parameter of interest, 95% credible intervals and the posterior probability that a coefficient parameter β is bigger than zero $P(\beta > 0)$. A 95% credible interval demarcates the range of values that comprise 95% of probability mass of our posterior beliefs, such that no value inside the CI has a higher probability than any point outside (see, for example, E. T. Jaynes & Kempthorne, 1976; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). We judge there to be 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]

[4]Note that we preregistered an analysis within the frequentist framework. However, due to severe convergence issues with complex random effect structures, we were not able to run the desired 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 we decided to back up the preregistered analysis with the conceptually desired random effect structure in the present Bayesian analysis. This more conservative approach resulted in the same overall results. Both analyses and their results can be assessed in our R scripts on our osf repository.

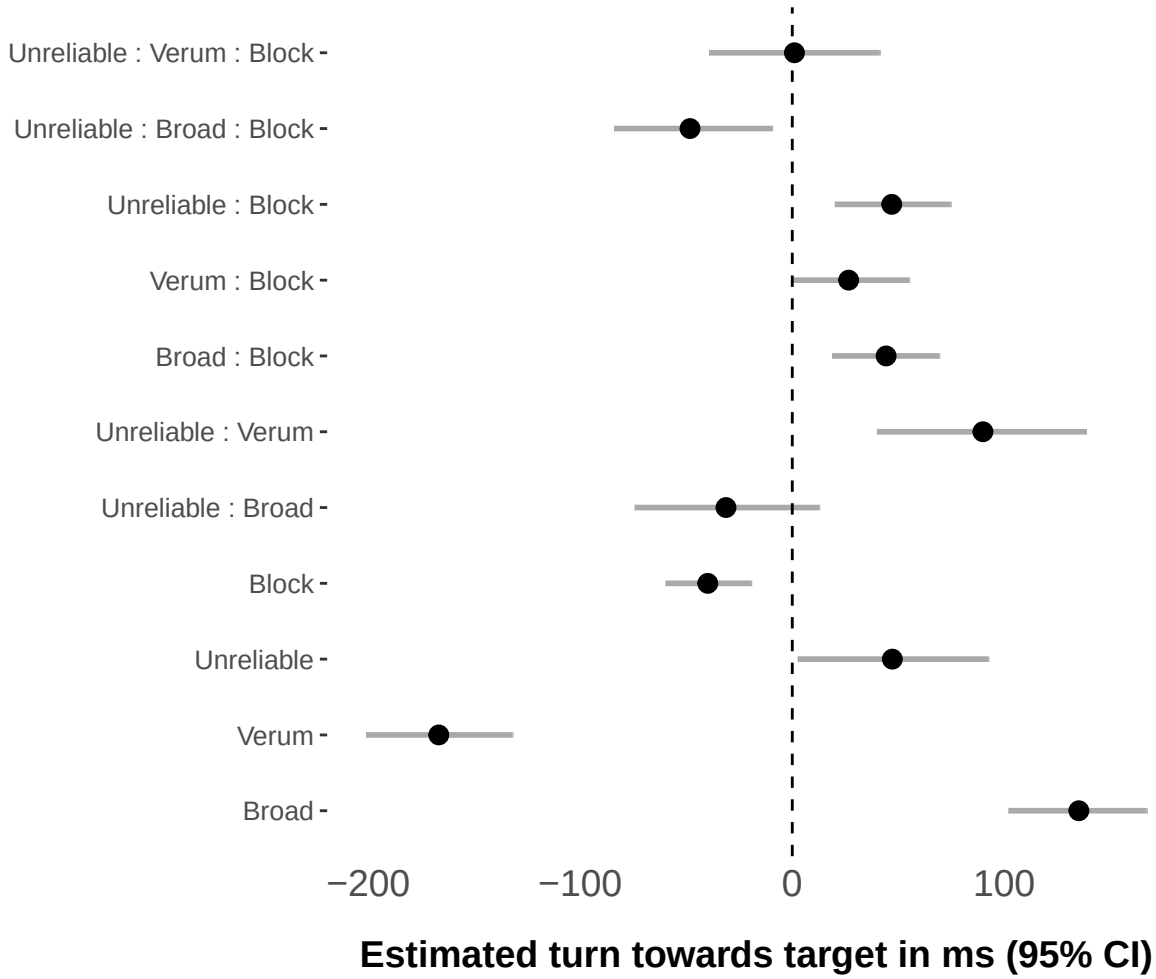

Figure 3. Forest plot of the estimated turn-towards-target measurement. A positive difference indicates evidence for a positive adjustment of the intercept which is contrastive focus in the reliable group in the middle of the experiment (scaled block = 0). Horizontal lines represent 95% credible intervals.

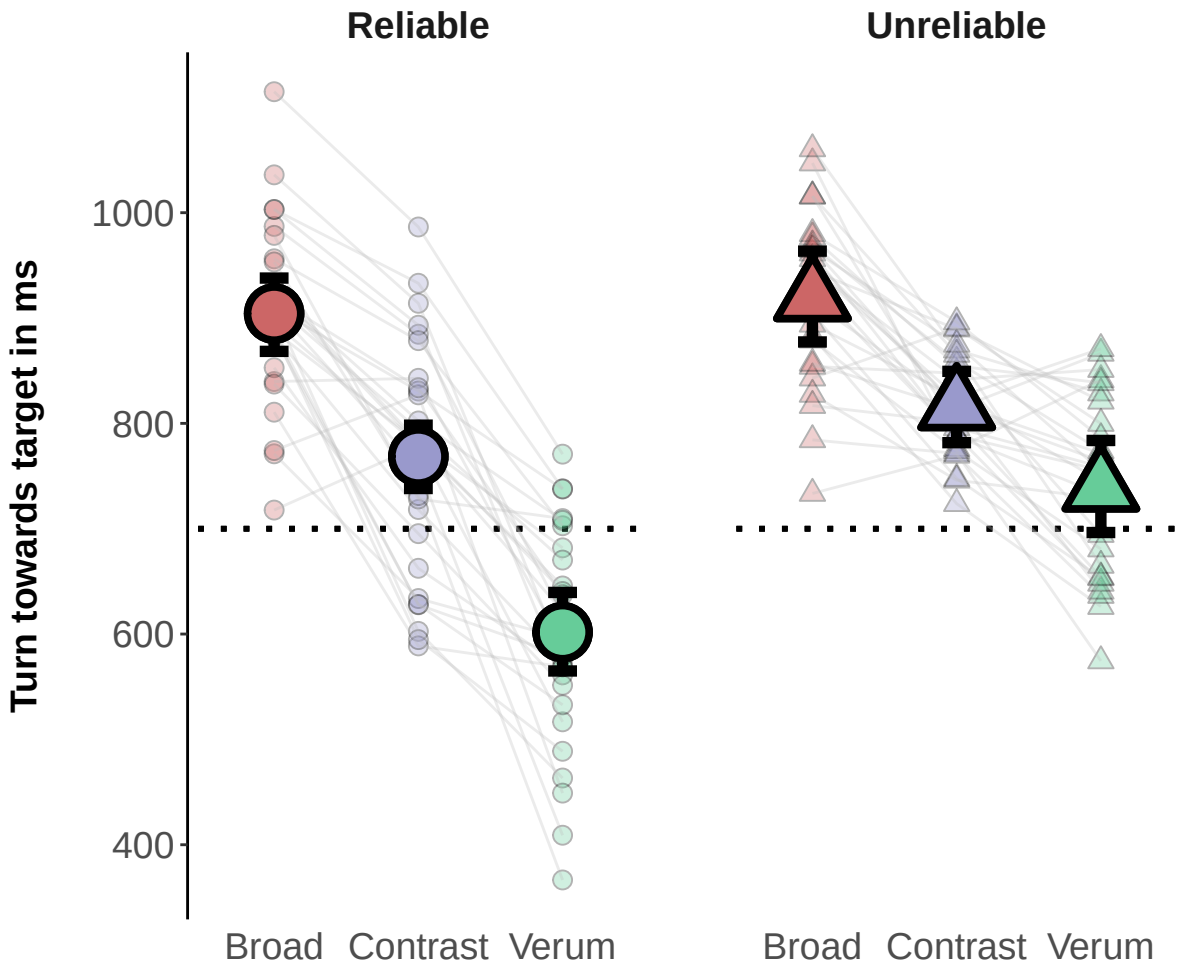

Figure 4. Estimates and 95% credible intervals for the turn-towards-target measurement across focus conditions and listener groups. Semi-transparent small points are average values for each subject. Solid grey lines group individual subject's values across focus condition. The dotted line indicates the average acoustic onset of the referent.

Since both RT and TTT exhibit similar patterns, we will report both measurements together. We used dummy coding with contrast focus in the RS group as the baseline. The overall coefficients and corresponding 95% CIs can be visually inspected in Figure 3 and Figure B1 in the appendix. Figure 4 visualizes the comparison between conditions for TTT. There is substantial support for verum focus eliciting faster responses than contrastive focus (RT: $\hat{\beta} = -147$, 95% CI $= [-169, -125]$, $P(\beta > 0) \approx 0$; TTT: $\hat{\beta} = -167$, 95% CI $= [-201, -132]$, $P(\beta > 0) \approx 0$). Moreover, contrastive focus elicits faster responses than broad focus (RT: $\hat{\beta} = 140$, 95% CI $= [113, 166]$, $P(\beta > 0) \approx 1$; TTT: $\hat{\beta} = 135$, 95% CI $= [102, 167]$, $P(\beta > 0) \approx 1$).

These results indicate that TTT appears to be a sensitive measurement. Lexical disambiguation in the broad focus condition starts to manifest itself in movements towards the target at around 200 ms after the onset of the lexically disambiguating cue (starting on average at 700 ms). Taking this as a baseline, the intonationally informed responses exhibit significantly earlier TTTs reflecting the anticipatory nature of responses.

There is evidence that this anticipatory behaviour is modulated by unreliable exposure. Contrastive focus in the US group might elicit slower responses than in RS (RT: $\hat{\beta} = 78$, 95% CI $= [-5, 162]$, $P(\beta > 0) = 0.97$; TTT: $\hat{\beta} = 47$, 95% CI $= [3, 93]$, $P(\beta > 0) = 0.98$). However, evidence for this slow down is weak, in fact the CIs for RTs include zero and those for TTT only barely exclude it. In contrast, there is a clear modulation of this effect in verum focus which is strongly affected by GROUP, with a substantial response time decrease (RT: $\hat{\beta} = 89$, 95% CI $= [58, 120]$, $P(\beta > 0) \approx 1$; TTT: $\hat{\beta} = 90$, 95% CI $= [40, 139]$, $P(\beta > 0) \approx 1$).

These temporal effects changed dynamically across the course of the experiment (see Figure 5). In RS, subjects' responses to contrastive focus become quicker throughout the experiment. (RT: $\hat{\beta} = -63$, 95% CI $= [-82, -43]$, $P(\beta > 0) \approx 0$; TTT: $\hat{\beta} = -40$, 95% CI $= [-60, -19]$, $P(\beta > 0) \approx 0$). There is no strong evidence for a difference in the temporal facilitation effect between contrast and verum conditions, but posteriors over

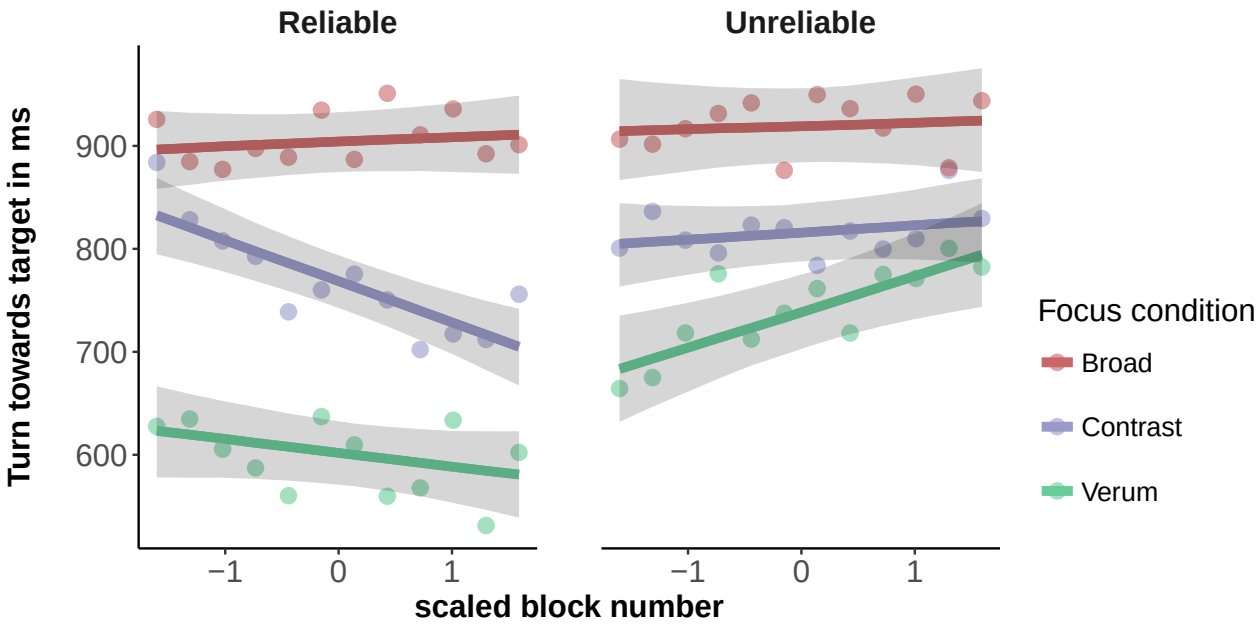

Figure 5. Estimated TTT values (lines) as a function of block number (scaled), dependent on focus condition and listener group. Shaded ribbons correspond to 95% credible intervals. Semi-transparent points correspond to average values for each block.

coefficients still suggest that it is not unlikely, given model and data, that exposure to reliable input more strongly accelerated disambiguation in the contrast condition over the course of the experiment (RT: $\hat{\beta} = 11$, 95% CI $= [-5, 27]$, $P(\beta > 0) = 0.91$; TTT: $\hat{\beta} = 27$, 95% CI $= [-1, 55]$, $P(\beta > 0) = 0.97$). In the broad focus condition the facilitation effect is clearly reduced (RT: $\hat{\beta} = 28$, 95% CI $= [14, 42]$, $P(\beta > 0) \approx 1$; TTT: $\hat{\beta} = 44$, 95% CI $= [19, 69]$, $P(\beta > 0) \approx 1$).

In the unreliable group, listeners become slower over time when responding to contrastive focus (RT: $\hat{\beta} = 59$, 95% CI $= [30, 88]$, $P(\beta > 0) \approx 1$; TTT: $\hat{\beta} = 47$, 95% CI $= [20, 75]$, $P(\beta > 0) \approx 1$). This slowing down is reduced in broad focus, in which listeners' response times stagnated throughout the experiment (RT: $\hat{\beta} = -34$, 95% CI $= [-56, -13]$, $P(\beta > 0) \approx 0$; TTT: $\hat{\beta} = -48$, 95% CI $= [-84, -9]$, $P(\beta > 0) = 0.01$). Although the mean predictions in Figure 5 visually suggest a tendential difference, there is no indication to believe that verum focus had a different dynamic profile from

contrastive focus (RT: $\hat{\beta} = 11$, 95% CI $= [-13, 36]$, $P(\beta > 0) = 0.82$; TTT: $\hat{\beta} = 1$, 95% CI $= [-39, 41]$, $P(\beta > 0) = 0.51$).

4. Discussion

4.1 Summary of results

The present data suggest that intonational information, if used reliably according to the conventions of the respective speech community, can facilitate intention recognition in the presence of relevant discourse information (e.g., *inter alia*, Dahan et al., 2002; Ito & Speer, 2008; Kurumada et al., 2014a; Roettger & Stoeber, 2017; Watson et al., 2008; Weber et al., 2006). The acoustically early cue associated with verum focus allows listeners to infer the intended referent long before the lexical material becomes available. Beyond that, listeners also use the absence of this cue (no accent on the auxiliary) to anticipate the contrastive interpretation. This inference does not happen as fast as in the verum focus condition but happens still earlier than lexical disambiguation (broad > contrastive > verum). The findings that the absence of a pitch accent can be taken as a weak cue for intention recognition is in line with previous experimental work (see Carbary et al., 2015; Weber et al., 2006) but runs counter to recent observations by Kurumada et al. (2014a) and Kurumada et al. (2014b). We will come back to this issue below, where we argue that these findings are compatible with rational immediate cue integration.

Intonational cue exploitation depends on the estimated reliability of form-function mappings. Listeners appear to weigh down the informational value of intonation in the unreliable group, but still exploit intonational cues to some degree. Our data further suggest that incremental cue exploitation changes dynamically throughout exposure. Exposure to reliable cues leads to earlier cue integration in the contrast and verum condition, likely with a more pronounced acceleration for contrastive focus. Exposure to unreliable cues leads to slower decisions throughout the experiment. However, despite these dynamic changes and their differences between exposure groups, listeners

exploited the absence of a pitch accent already at the earliest stages of the experiment in either exposure group. This suggests that intonational cue exploitation can be differentially facilitated or modulated by exposure, but is not likely just a mere task-adaptation or experimental artifact, but rather part of the natural disposition of language users.

4.2 Rational predictive processing

For a rational predictive interpreter (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016), there should not necessarily be a categorical difference between presence and absence of a cue (Hsu et al., 2017). What matters are differences in the listeners' beliefs about how likely a speaker is to produce a particular contour to convey a particular meaning. The question therefore arises whether the following patterns observed in our data are compatible with rapid rational cue integration:

Obs 1: decision making in the contrast condition ("absence of an early cue") is slower than in the verum condition ("presence of an early cue");

Obs 2: exposure to reliable input possibly speeds up decisions in the contrast condition more than in the verum condition, while no apparent difference between conditions shows under unreliable input.

The following explores a model of rapid rational cue integration in terms of Bayesian inference to explain these observations.[5] Observation 1 follows from natural and plausible assumptions about asymmetries in the relevant production likelihoods. Observation 2 requires additional assumptions about belief dynamics (how listeners adapt their beliefs about speaker production during the experiment) and the link function between listener beliefs and the TTT measure. In the absence of clarity about

[5]The modeling is entirely post-hoc, conceived and fine-tuned after the data was known. Accordingly, it should be treated as entirely exploratory.

these two aspects, our model is at best preliminary and therefore supports only the modest, but still important conclusion that observation 2 is compatible with *some* rational analysis of rapid cue integration, belief dynamics and link function.

We consider a rational Bayesian listener who reasons about the speaker's likelihood of producing different messages (sentences with particular intonational properties) to convey different meanings (referents). 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 probability of the target is higher earlier during the sentence: the more certain a participant is at the current stage, the more likely it is that she makes a decision already and turns towards the target. This assumption is in line with the general idea of ballistic accumulator models (e.g., Ratcliff & McKoon, 2008), namely that evidence in favor of a choice or hypothesis accumulates incrementally and results in execution if a critical mass is met.

It suffices to consider cases with a preceding discourse question "Did the wuggy pick up referent r_g?" (with r_g the *given referent* and r_c the *competitor*), and to consider two partial utterances of "The Wuggy has": one where "has" bears a pitch accent, as in the *verum* condition; another where it does not, as in the *contrast* condition. If we write V for the former and C for the latter, a rational comprehender's posterior odds in favor of the target referent after observing either utterance is given by:

$$\frac{P(r_g \mid V)}{P(r_c \mid V)} = \frac{P(V \mid r_g)}{P(V \mid r_c)} \; \frac{P(r_g)}{P(r_c)}$$
$$\frac{P(r_c \mid C)}{P(r_g \mid C)} = \frac{P(C \mid r_c)}{P(C \mid r_g)} \; \frac{P(r_c)}{P(r_g)}$$

According to our link hypothesis, the TTT measure is a strictly decreasing function of the posterior odds in favor of the target referent. Although the prior odds also affect the posterior odds, the impact of the likelihood ratio is in focus, if the prior odds are reasonably close to 1.[6] The relevant likelihoods can be parameterized as in the following table (where $v^{|r_g} = P(V \mid r_g)$ etc.):

[6]This assumption is supported by the observation that any strong prior biases should influence mouse

	V	C
r_g	$v^{\mid r_g}$	$c^{\mid r_g}$
r_c	$v^{\mid r_c}$	$c^{\mid r_c}$

While any particular choice of numbers for production likelihoods would be arbitrary, a number of constraints are rather uncontroversial: (i) pitch accent on the auxiliary is unlikely when the speaker wants to refer to the competitor referent ($v^{\mid r_c} < c^{\mid r_c}$); (ii) pitch accent is more likely for r_g than for r_c ($v^{\mid r_g} > v^{\mid r_c}$); (iii) verum focus is overall less frequent ($v^{\mid r_g} + v^{\mid r_c} < c^{\mid r_g} + c^{\mid r_c}$). Based on these general constraints Appendix A proves the following

Proposition 1. *Presence of a pitch accent provides more observational evidence in favor of the target r_g than absence provides in favor of the target r_c:* $\frac{v^{\mid r_g}}{v^{\mid r_c}} > \frac{c^{\mid r_c}}{c^{\mid r_g}}$.

This explains Observation 1. Rather than assuming a categorical difference between "absence" and "presence" of a cue, the rational cue integration model traces a different evidential value back to a difference in overall production frequencies. To see this, consider concrete numbers for illustration:

	V	C
r_g	.6	.4
r_c	.1	.9

The difference between relevant production likelihoods is identical:

$$v^{\mid r_g} - v^{\mid r_c} = .6 - .1 = 0.5 = .9 - .4 = c^{\mid r_c} - c^{\mid r_g}$$

trajectories already at the earliest positions in the sentence, which is not seen in the data. Moreover, heavily skewed prior odds are also not supported by frequency: in our design the given referent and the competitor occurred equally often as the target mentioned in the given sentence.

But the same difference in likelihoods yields a higher probability ratio for the lower probability event V:

$$\frac{v^{|r_g}}{v^{|r_c}} = \frac{.6}{.1} = 6 > 2.25 = \frac{.9}{.4} = \frac{c^{|r_c}}{c^{|r_g}}$$

To explain Observation 2 further assumptions are required about belief dynamics and the link function between posterior evidence and TTT. A simple model of belief dynamics assumes that comprehenders keep track of non-normalized scores (e.g., a count of the number of times in which they recently observed speakers use a certain form-meaning pair). For instance, we might consider:

	V	C
r_g	60	40
r_c	10	90

Beliefs about speaker production probabilities are derived from these scores by normalization in the usual way. If listeners observe a co-occurrence of an intonational pattern (V or C) with a referent (r_g or r_c), they increment the relevant score by one. Consequently, each block in the reliable group will increment $v^{|r_g}$ and $c^{|r_c}$ by 2, since there are two trials each of reliable verum and contrast conditions per block. In the unreliable condition there are additional unreliable trials in each block, one where V is paired with r_c and one where C is paired with r_g. We therefore increment counts by one for these pairs for each block in the unreliable condition.

As for the mapping from posterior odds to TTT, the latter must have a finite lower bound to which it converges from above as posterior odds grow to infinity. An exponential decay function is a natural choice:

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

The resulting TTT dynamics over 12 blocks of reliable or unreliable trials for the numeric example given above is plotted in Figure 6. The model captures the qualitative pattern

that we wanted to explain, namely that reliable input mainly speeds up responses in the contrast condition, while no marked contrast is seen under unreliable input. We conclude that superficially surprising patterns in our data are compatible with *some* model of rational belief dynamics and rapid cue integration. More data on adaptation to (partially) reliable intonational cues and their effect on manual response dynamics would be necessary to formulate a definite model with more confidence.

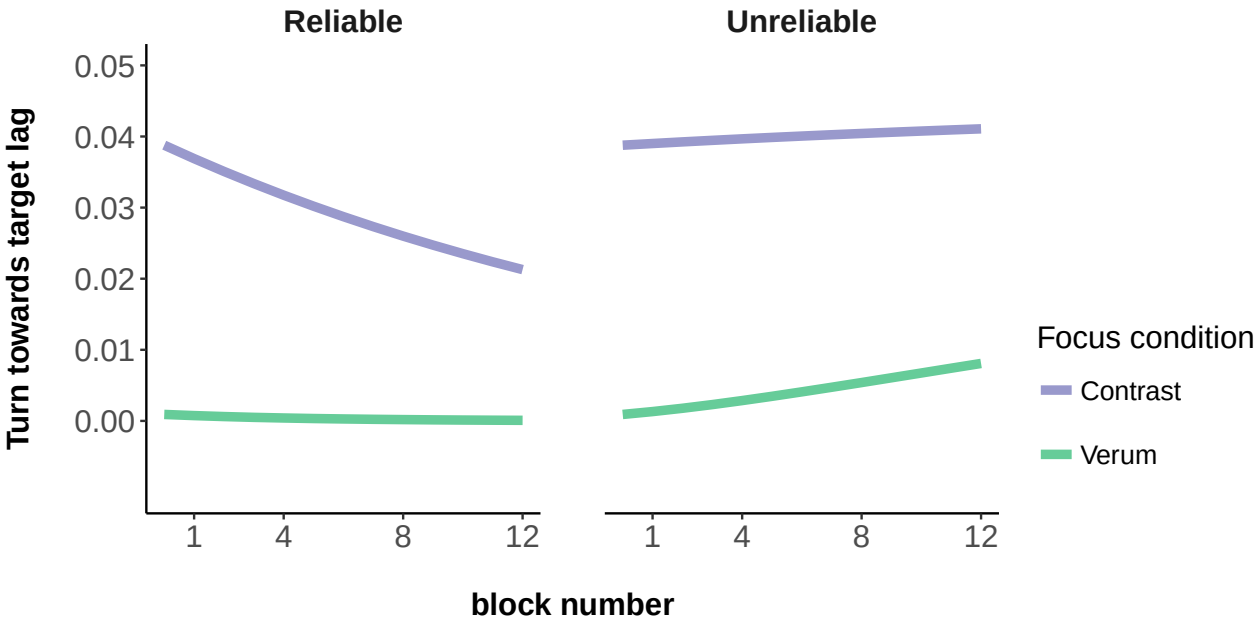

Figure 6. Predictions of a model of belief dynamics and rational rapid cue integration with an exponential link function between posterior odds and the TTT measurement.

Conclusion

Intonational information can provide early cues to speaker-intended information, even before lexically disambiguating information is available. We presented data that suggest that listeners are able to rapidly exploit intonational cues during online processing. As demonstrated by the time at which participants started to move their mouse consistently towards the final target (the turn-towards-the-target (TTT) measure), 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-present or discourse-new referent. By manipulating the stochastic reliability of intonational cues in a between-subject design, we further found that listeners seem to generally anticipate a reliable mapping, starting to exploit 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 exploit intonation predictively. Over the course of the experiment consistent reliable input leads to earlier turns towards the target, more markedly for one type of intonational cue (the contrast condition with absence of a pitch accent on the early auxiliary), while partly unreliable input impeded exploitation of presence and absence of an intonational cue in similar ways. We presented an exploratory, post-hoc 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 exploit intonational information and update their expectations about speaker production likelihoods dynamically.

Appendix A

Proof of proposition

Five assumptions are necessary to prove Proposition 1, three of which were mentioned in the main text already and another two of a more technical character.

Ass 1: pitch accent on "has" is unlikely when the speaker wants to refer to the competitor referent ($v^{|r_c} < c^{|r_c}$)

Ass 2: pitch accent is more likely for r_g than for r_c ($v^{|r_g} > v^{|r_c}$)

Ass 3: verum focus is overall less frequent ($v^{|r_g} + v^{|r_c} < c^{|r_g} + c^{|r_c}$)

Ass 4: there are no other relevant realizations of "has" beyond V and C in the microcosm of the experiment ($v^{|r_c} = (1 - c^{|r_c})$ and ($v^{|r_g} = (1 - c^{|r_g})$)[7]

Ass 5: all production likelihoods are positive ($v^{|r_g}, c^{|r_g}, v^{|r_c}, c^{|r_c} > 0$)

We establish three helpful results first.

Corollary 1. $c^{|r_c} > v^{|r_g}$

Proof.

$$v^{|r_g} + v^{|r_c} < c^{|r_g} + c^{|r_c} \qquad \text{[by Ass 3]}$$

$$v^{|r_g} + (1 - c^{|r_c}) < (1 - v^{|r_g}) + c^{|r_c} \qquad \text{[by Ass 4]}$$

$$2v^{|r_g} - 1 < 2c^{|r_c} - 1$$

$$v^{|r_g} < c^{|r_c}$$

□

Corollary 2. $v^{|r_g} - v^{|r_c} = c^{|r_c} - c^{|r_g}$

[7] A weaker assumption would give the same result, namely that there are no other relevant realizations of "has" beyond V and C that would be substantially more likely for one referent than for the other.

Proof.

$$
\begin{aligned}
v^{|r_g} - v^{|r_c} &= c^{|r_c} - c^{|r_g} \\
v^{|r_g} - v^{|r_c} &= (1 - v^{|r_c}) - (1 - v^{|r_g}) && \text{[by Ass 4]} \\
v^{|r_g} - v^{|r_c} &= v^{|r_g} - v^{|r_c}
\end{aligned}
$$

□

Fact 1. *Function* $f(x) = \frac{x+c}{x}$ *with* $x, c > 0$ *is strictly monotone decreasing and concave.*

Proof. 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 $x, c > 0$. □

With these in place we can proof Proposition 1 as follows.

Proof of Theorem 1. We need to show that:

$$\frac{v^{|r_g}}{v^{|r_c}} > \frac{c^{|r_c}}{c^{|r_g}}.$$

By Corollary 2 and Assumption 2 we can rewrite this as:

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

By Corollary 1 we know that $x < x'$. The result follows from Fact 1 and assumption 5.

□

Appendix B

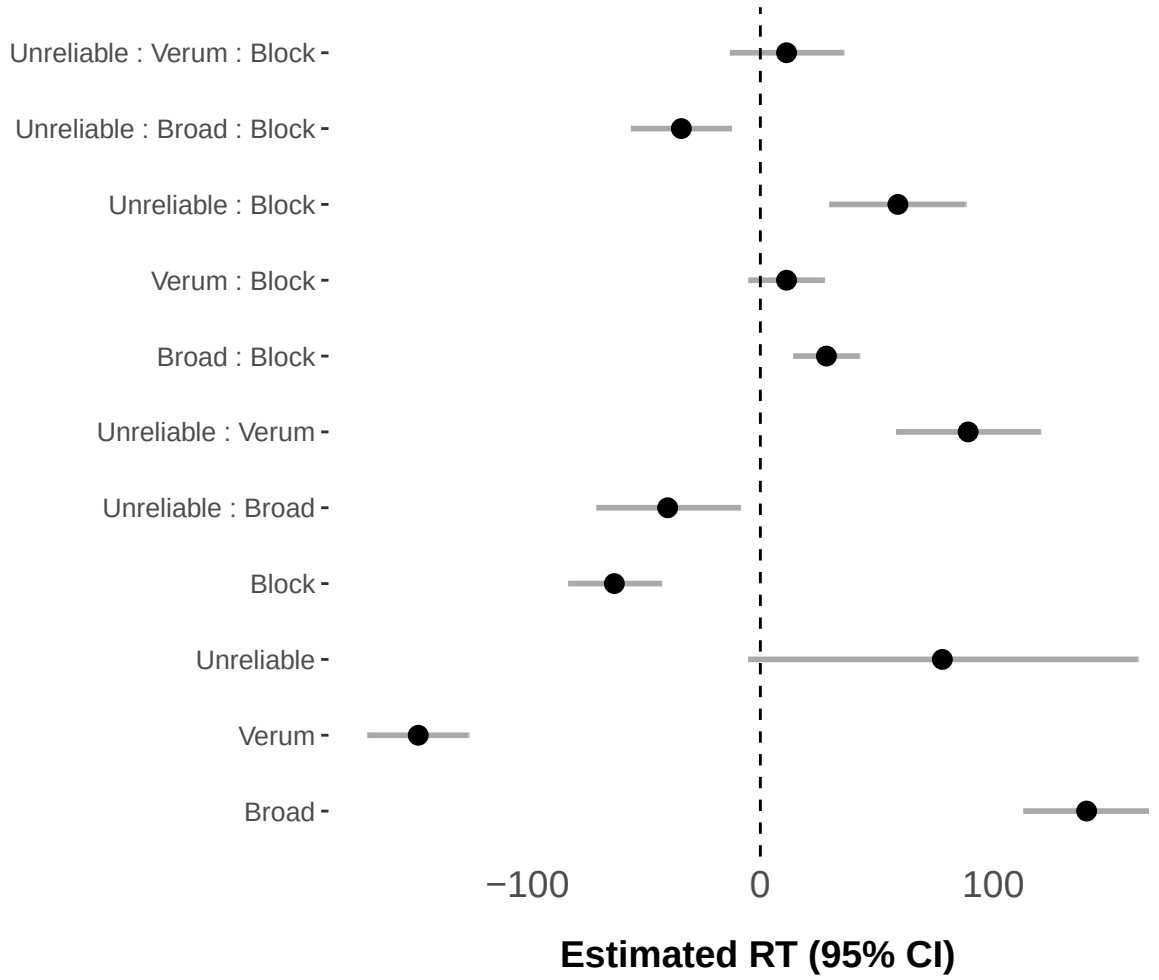

Figure B1. Forest plot of the estimated reaction times. A positive difference indicates evidence for a positive adjustment of the intercept which is contrastive focus in the reliable group in the middle of the experiment (scaled block = 0). Horizontal lines represent 95% credible interval.

Appendix C

References

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

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.

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

Büring, D. (2009). Towards a typology of focus realization. In M. Zimmermann & C. Féry (Eds.), *Information structure* (pp. 177–205). Oxford: Oxford University Press.

Carbary, K., Brown, M., Gunlogson, C., McDonough, J. M., Fazlipour, A., & Tanenhaus, M. K. (2015). Anticipatory deaccenting in language comprehension. *Language, Cognition and Neuroscience*, *30*(1-2), 197–211.

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.

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.

Dennison, H. Y., & Schafer, A. J. (2010). Online construction of implicature through contrastive prosody. In *Speech prosody 2010-fifth international conference*.

Dotan, D., Meyniel, F., & Dehaene, S. (2018). On-line confidence monitoring during decision making. *Cognition*, *171*, 112–121. doi:https://doi.org/10.1016/j.cognition.2017.11.001

Féry, C., & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents

in german. *Journal of Phonetics*, *36*(4), 680–703.

Fine, A. B., & Florian Jaeger, T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*(3), 578–591.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998. doi:10.1126/science.1218633

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift Für Sprachwissenschaft*, *35*(1), 3–44. doi:10.1515/zfs-2016-0002

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.

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. doi:10.1016/j.tics.2016.08.005Au

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., & Sedivy, J. C. (2011). 10 the effect of speaker-specific information on pragmatic inferences. *The Processing and Acquisition of Reference*, 239.

Hsu, A. S., Horng, A., Griffiths, T. L., & Chater, N. (2017). When absence of evidence is evidence of absence: Rational inferences from absent data. *Cognitive Science*, *4*, 1155–1167. doi:10.1111/cogs.12356

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.

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). Dordrecht: Springer Netherlands. doi:10.1007/978-94-010-1436-6_6

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

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

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.

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., & Tanenhaus, M. (2014b). Rapid adaptation in online pragmatic interpretation of contrastive prosody. In *Proceedings of the cognitive science society* (Vol. 36).

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

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, *23*(1), 103–123. doi:10.3758/s13423-015-0947-8

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

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

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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* (Vol. 3). Language Science Press.

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 CogSci 39* (pp. 3010–3015). Austin, TX: Cognitive Science Society.

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.

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

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.

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.

Supplementary file II - Acoustic resynthesis of stimuli

Timo B. Roettger[1,2] & Michael Franke[3]

[1] University of Cologne

[2] Northwestern University

[3] University of Tübingen

Author Note

Correspondence concerning this article should be addressed to Timo B. Roettger, Herbert-Lewin-Str. 6, D-50931 Cologne. E-mail: timo.b.roettger@gmail.com

Supplementary file II - Acoustic resynthesis of stimuli

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 FOCUS conditions exhibit the same temporal characteristics for each sentence (i.e. all lexical information of the referent becomes available at the same time), sentences were manipulated and resynthesized using Boersma and Weenink (2016) applying the following procedure.

Step 1 - baseline: We took the original stimuli produced with a verum focus intonation as a departure point because they can easily be resynthesized into prosodic patterns corresponding to contrastive or broad focus, without creating mismatching acoustic information. We took one prototypical statement produced with verum focus and isolated the first part of the sentence ("Der Wuggy hat" "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" (i.e. "dann die Birne aufgesammelt" "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 twelve different base sentences, exhibiting the same temporal landmarks up to "hat". The concatenated stimuli do not contain any auditory residuals of the splicing procedure.

Step 2: In the next step, we manipulated the duration of "hat" and the stressed syllable of each referent (e.g. "BIRne" "pear"). Too 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 the stimuli for the verum focus condition and were further processed for the manipulation of broad focus and contrastive focus.

For the broad focus condition, 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 broad focus in German (e.g. Grice, Ritter, Niemann, & Roettger, 2017).

For the contrastive focus 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 its stressed syllable. Following the rise, f_0 fell (50 Hz) towards the end of the stressed syllable of the referent. The rest of the utterance remained low, resulting in a rise-fall on the accented referent, commonly observed for contrastive focus in German (e.g. Grice et al., 2017).

See Figure 1 for a visual example and retrieve all audio stimuli from https://osf.io/dnbuk/.

References

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

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

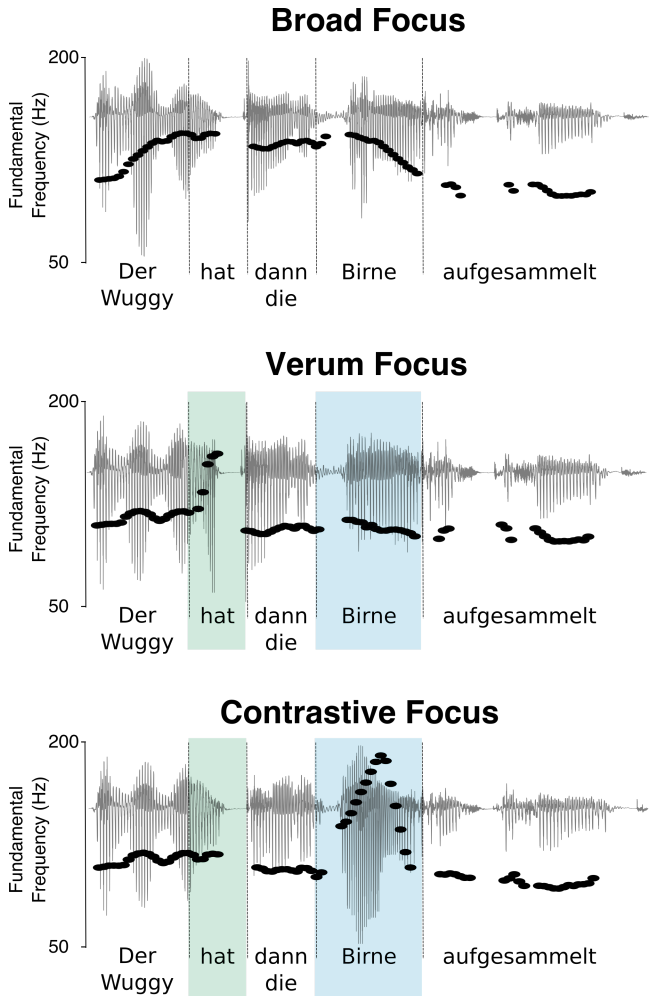

Figure 1. Representative waveforms and f_0 contours for broad, verum and lexical focus as resynthesised for the present experiment. Example trial corresponds to <Der Wuggy hat dann die Birne aufgesammelt.> 'The wuggy has picked up the pear then.' Lime green box indicates the auxiliary <hat>; light blue box indicates the referential expression.

Supplementary file I - Spatial analysis

Timo B. Roettger[1,2] & Michael Franke[3]

[1] University of Cologne

[2] Northwestern University

[3] University of Tübingen

Author Note

Correspondence concerning this article should be addressed to Timo B. Roettger, Herbert-Lewin-Str. 6, D-50931 Cologne. E-mail: timo.b.roettger@gmail.com

Supplementary file I - Spatial analysis

1 Descriptive assesment of trajectories

Figure 1 shows the averaged x,y-coordinates of time- and space-normalized trajectories for both listener groups. In the reliable listener group, verum focus turns towards the target relatively low on the y-axis, making a more direct pathway to the target. Broad focus first gravitates towards the horizontal mid point before curving to the target somewhat higher on the y-axis. Similarly, contrastive focus gravitates first towards the mid point, detours then somewhat towards the competitor and eventually turns towards the target. These patterns are strongly reduced in the unreliable group.

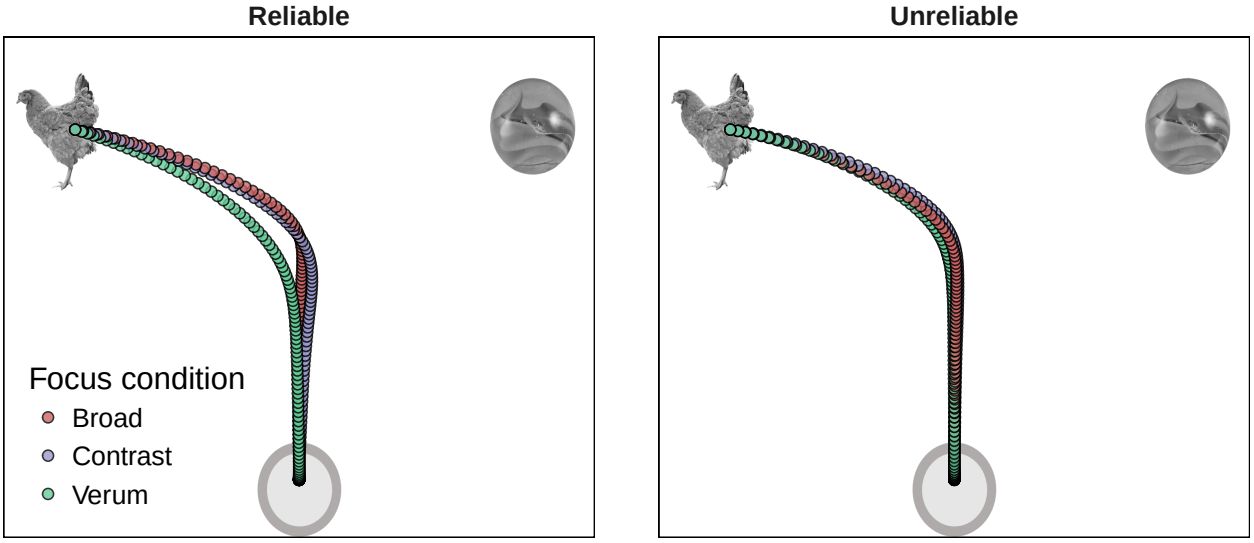

Figure 1. Averaged x,y-coordinates of time- and space-normalized trajectories for the reliable-speaker group and the unreliable-speaker group.

2 Inferential assesment

We report two spatial measures: (i) the area under the curve (AUC, as extracted via the "mt_derivatives()" function), operationalized by the geometric area between the observed trajectory and an idealized straight-line trajectory drawn from the start and end points (Freeman & Ambady, 2010); and (ii) the deviation away from the medial axis (Xneg, as

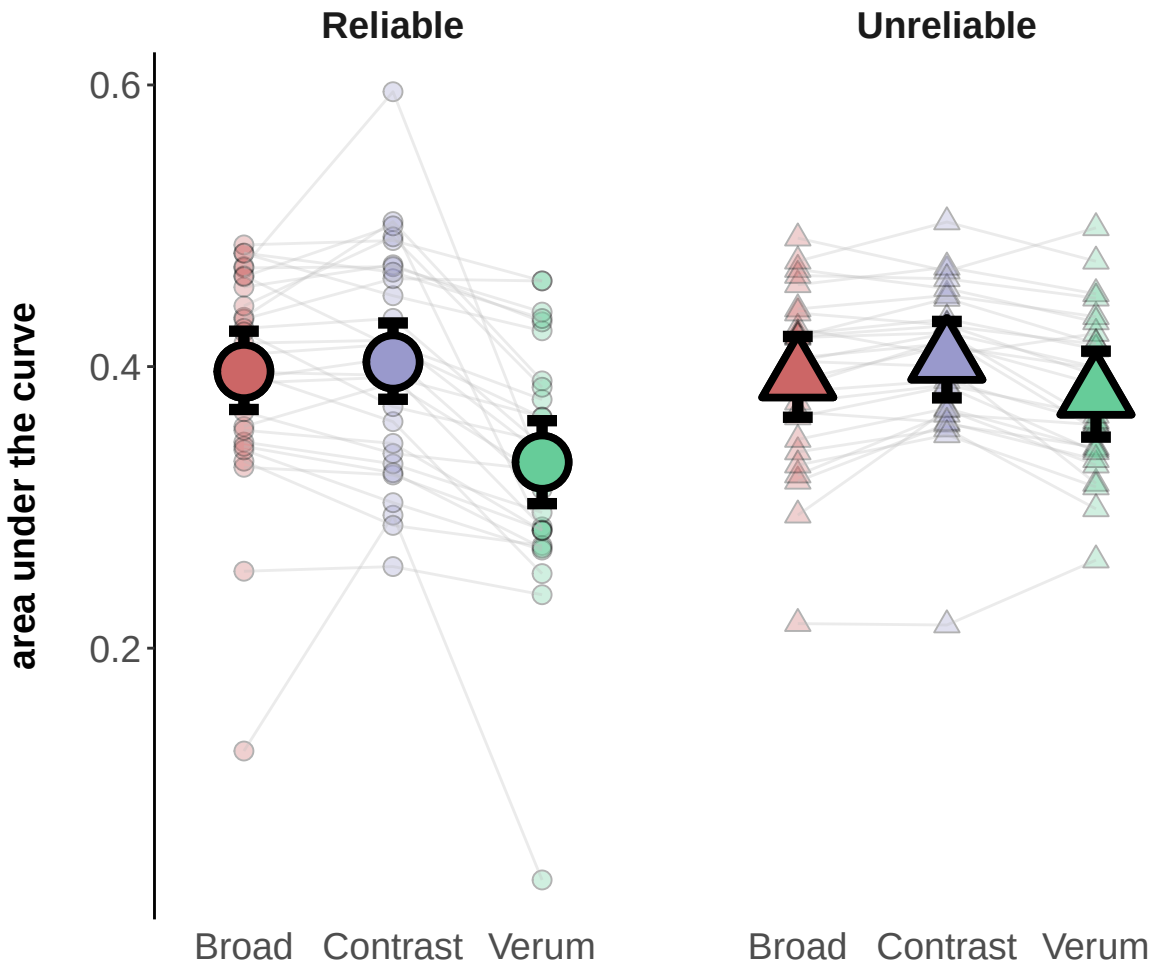

Figure 2. Posterior expected values and 95%-credible intervals for the Area-Under-the-Curve (AUC) measurement across focus conditions and listener groups. Semi-transparent small points are average values for each subject. Solid grey lines link individual subject's mean values across focus condition.

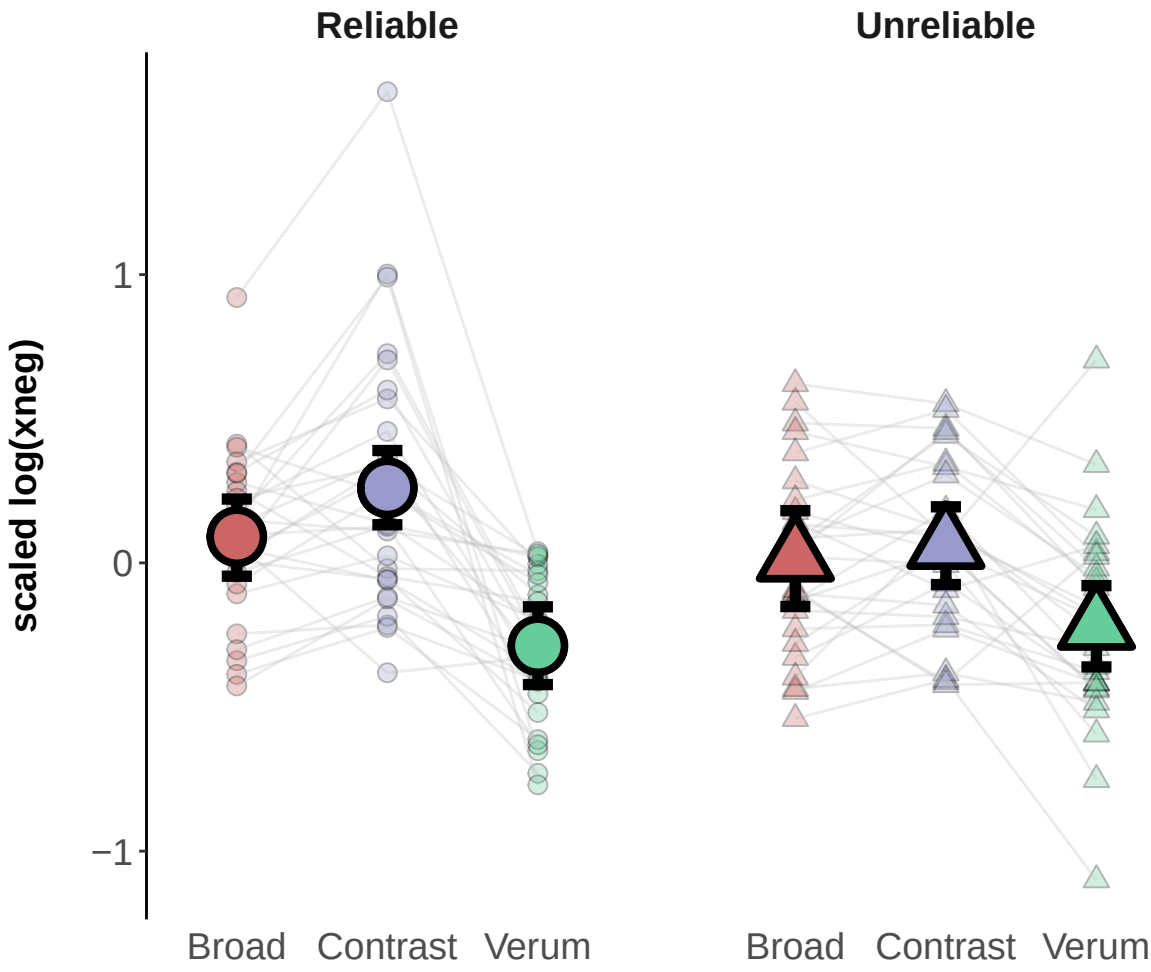

Figure 3. Posterior expected values and 95%-credible intervals for the scaled and log-transformed xneg measurement across focus conditions and listener groups. Semi-transparent small points are average values for each subject. Solid grey lines link individual subject's mean values across focus condition.

extracted via the "mt_derivatives()" function) as an additional indicator of the competitor attraction. We fitted Bayesian hierarchical linear models to area-under-the-curve (AUC) and scaled, log-transformed xneg measurements as a function of FOCUS, GROUP and BLOCK and their interaction, using the Stan modelling language (Carpenter et al., 2016) and the package *brms* (Buerkner, 2016) in (R Core Team, 2017). We used weakly informative Gaussian priors centered around zero with $\sigma = 0.25$ (AUC) and $\sigma = 0.5$ (xneg), respectively, for all population-level regression coefficients. We used dummy coding with contrast focus in the RS group as the baseline, we will therefore report on the respective main effects and interactions in comparison to the baseline (see Figures 2, 3, 6 and 7. R-scripts and raw data are available in our osf repository.

There is substantial support for verum focus eliciting less AUC than contrastive focus in the reliable speaker group ($\hat{\beta} = -0.071$, 95% CI $= [-0.092, -0.051]$, $P(\beta > 0) \approx 0$). Contrastive and broad focus did not elicit different AUC values ($\hat{\beta} = -0.007$, 95% CI $= [-0.021, 0.007]$, $P(\beta > 0) = 0.17$).

There is evidence that this spatial asymmetry is modulated by unreliable exposure. Although, contrastive focus in the US group did not differ from the RS group ($\hat{\beta} = 0.002$, 95% CI $= [-0.036, 0.04]$, $P(\beta > 0) = 0.54$), there was a clear interaction with verum focus, such that unreliable exposure led to a decrease in AUC for verum focus ($\hat{\beta} = 0.046$, 95% CI $= [0.019, 0.076]$, $P(\beta > 0) \approx 1$) (see Figure 2).

For the xneg measurement, we find that contrastive focus elicited larger xneg values than both verum focus ($\hat{\beta} = -0.548$, 95% CI $= [-0.69, -0.405]$, $P(\beta > 0) \approx 0$), and broad focus ($\hat{\beta} = -0.17$, 95% CI $= [-0.285, -0.059]$, $P(\beta > 0) \approx 0$), with verum focus eliciting the lowest xnegs.

There is evidence that the observed spatial asymmetry in RS is modulated by unreliable exposure. Contrastive focus in the US group elicited lower xnegs than in the RS group ($\hat{\beta} = -0.199$, 95% CI $= [-0.38, -0.025]$, $P(\beta > 0) = 0.01$). There was also an interaction with verum focus, such that unreliable exposure led to an increase in xneg for

verum focus ($\hat{\beta} = 0.267$, 95% CI $= [0.081, 0.46]$, $P(\beta > 0) \approx 1$) (see Figure 3).

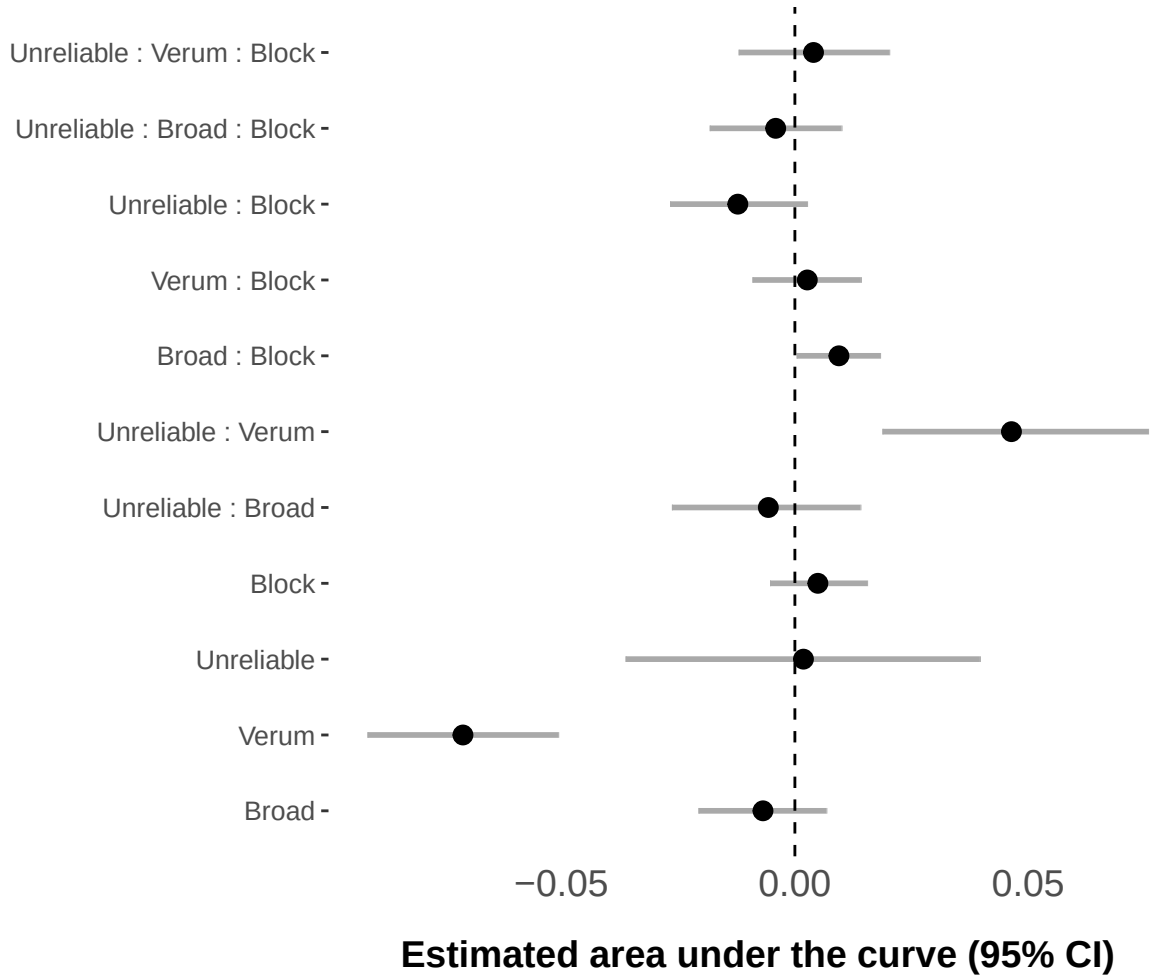

Figure 4. Forest plot of the estimated area under the curve measurement. A positive difference indicates evidence for a positive adjustment of the intercept which is contrastive focus in the reliable group in the middle of the experiment (scaled block = 0). Horizontal lines represent 95% credible interval.

There was almost no evidence that these spatial asymmetries changed dynamically across the course of the experiment (see Figures 6 and 7).

In the reliable-speaker group, subjects' mouse movements in the contrast focus condition do not appear to become more direct throughout the experiment (AUC: $\hat{\beta} = 0.005$, 95% CI $= [-0.005, 0.015]$, $P(\beta > 0) = 0.83$; xneg: $\hat{\beta} = -0.005$, 95% CI $= [-0.083, 0.072]$, $P(\beta > 0) = 0.45$). No evidence shows for a difference between contrast

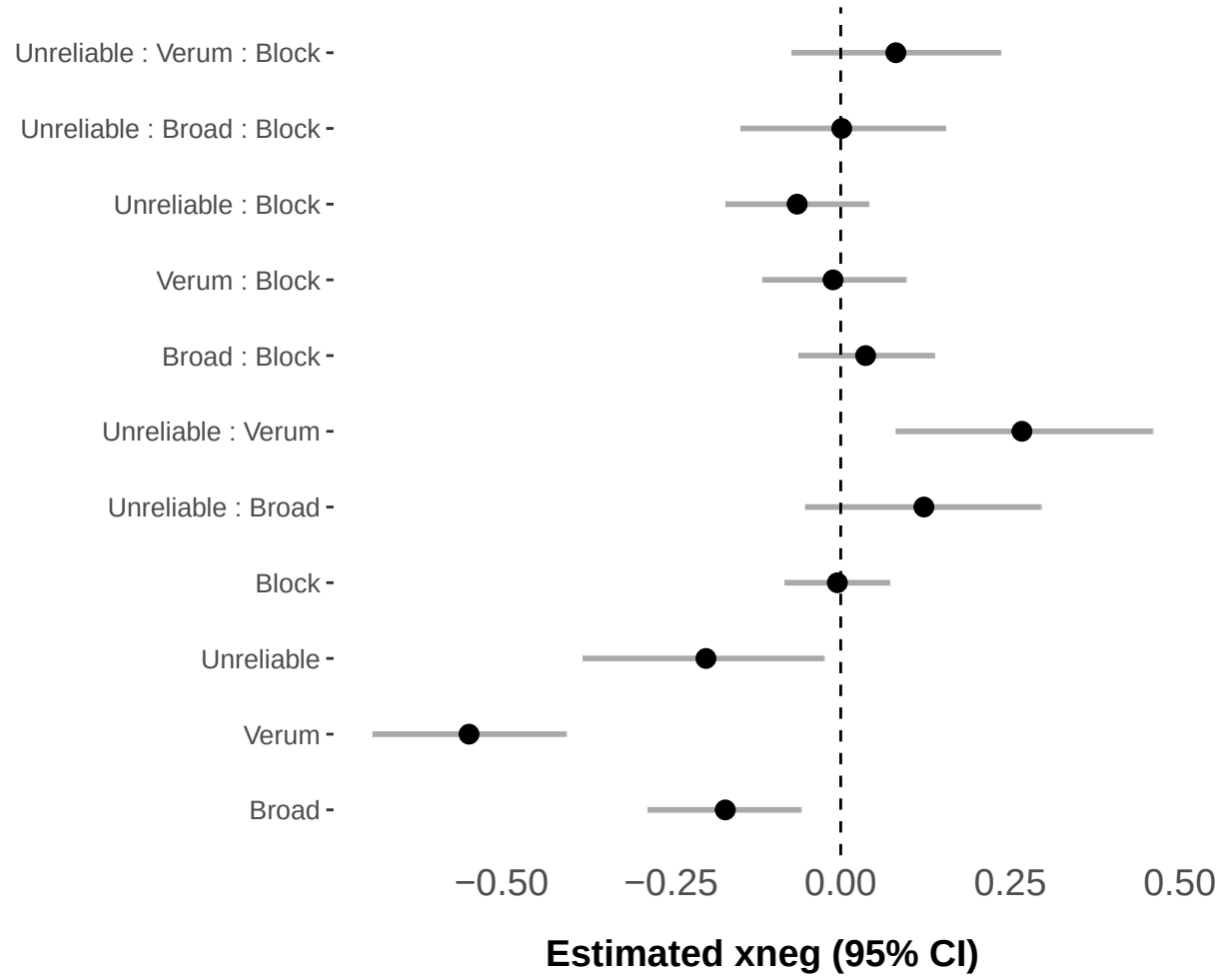

Figure 5. Forest plot of the estimated xneg measurement. A positive difference indicates evidence for a positive adjustment of the intercept which is contrastive focus in the reliable group in the middle of the experiment (scaled block = 0). Horizontal lines represent 95% credible interval.

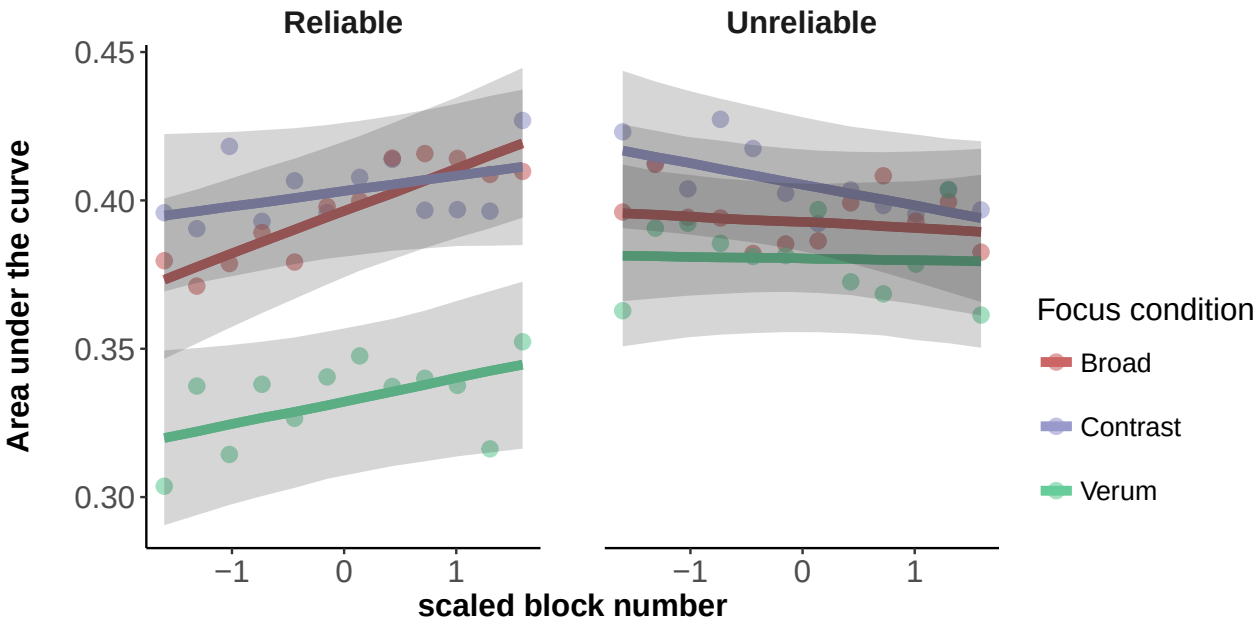

Figure 6. Estimated AUC values (lines) as a function of block number (scaled), dependent on FOCUS condition and reliability group. Shaded ribbons correspond to 95% credible intervals.

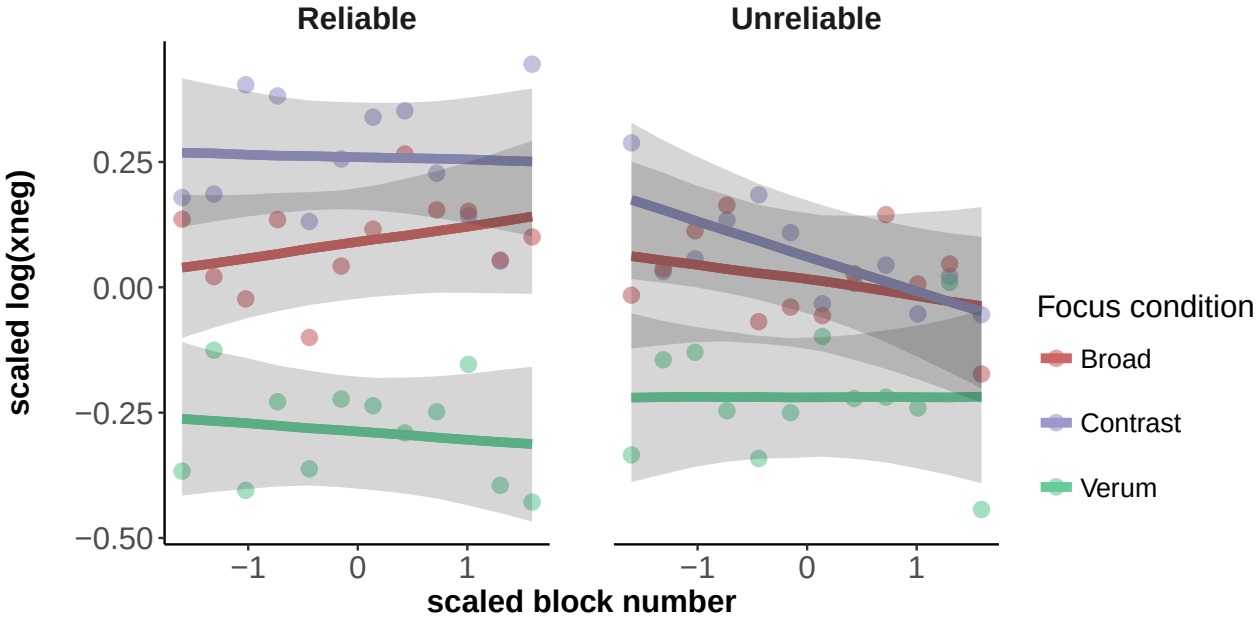

Figure 7. Estimated xneg values (lines) as a function of block number (scaled), dependent on FOCUS condition and reliability group. Shaded ribbons correspond to 95% credible intervals.

and verum conditions (AUC: $\hat{\beta} = 0.003$, 95% CI $= [-0.009, 0.014]$, $P(\beta > 0) = 0.68$; xneg: $\hat{\beta} = -0.011$, 95% CI $= [-0.116, 0.096]$, $P(\beta > 0) = 0.42$). In the broad focus condition, we see mild evidence for more efficient and direct mouse movements towards the target developing over the time course of the experiment (AUC: $\hat{\beta} = 0.009$, 95% CI $= [0, 0.018]$, $P(\beta > 0) = 0.98$; xneg: $\hat{\beta} = 0.037$, 95% CI $= [-0.062, 0.138]$, $P(\beta > 0) = 0.76$).

In the unreliable-speaker group, a potential tendency towards a temporal development of more direct movements over the course of the experiment showed in the contrast focus condition (AUC: $\hat{\beta} = -0.012$, 95% CI $= [-0.027, 0.003]$, $P(\beta > 0) = 0.06$; xneg: $\hat{\beta} = -0.064$, 95% CI $= [-0.17, 0.041]$, $P(\beta > 0) = 0.12$). We see no strong evidence in support of a belief that broad focus differed from contrastive focus (AUC: $\hat{\beta} = -0.004$, 95% CI $= [-0.018, 0.01]$, $P(\beta > 0) = 0.28$; xneg: $\hat{\beta} = 0.002$, 95% CI $= [-0.148, 0.154]$, $P(\beta > 0) = 0.51$).

3 Discussion

As opposed to the temporal dynamics of mouse movements, the spatial dynamics exhibit substantially more variability, introducing a large amount of uncertainty regarding their interpretation. We observe a spatial asymmetry in the reliable group such that mouse trajectories in the verum focus are more direct (smaller AUC and xneg) than broad focus. Additionally, contrastive focus elicits larger deviations towards the competitor than broad focus (larger xneg). Taken together, these effects suggest a spatial bias towards a given referent in the reliable group (more direct pathways in verum focus and less direct pathways in contrastive focus). This asymmetry collapses in the unreliable group. We did not anticipate these spatial biases, we will therefore refrain from post-hoc speculations. As the relationship between spatial and temporal information of the mouse trajectory remains unclear, these results do not necessarily conflict with our temporal analysis offered in the main paper. A spatial bias does not necessary come with a temporal cost. The available evidence suggests that these spatial biases are constant over

the course of the experiment, i.e. they do not change in light of new exposure and might thus be less indicative of listeners' adaptation to new exposure.

References

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

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.

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.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/