Integration of contextual-pragmatic and phonetic information in speech perception: An eye-tracking study

Eszter Ronai[1,*], Yenan Sun[1], Alan C. L. Yu[1], and Ming Xiang[1]

[1]Department of Linguistics, The University of Chicago, USA
[*]Corresponding author: ronai@uchicago.edu

Abstract Pragmatic information, such as inferences regarding upcoming coreference, has been shown to influence phonetic perception (Rohde & Ettlinger, 2012). Pragmatic information, however, comes in many forms. Using a Visual World Paradigm, tracking listeners' categorical responses and the time-course of information integration via eye movements, we investigated whether and how a different kind of pragmatic information, the contrastive function of prenominal adjectives (Sedivy et al., 1999), can affect listeners' perception of voicing in initial plosives. Our results suggest that the pragmatic contrast inference did not affect the behavioral judgments on phonetic categorization, but it did have (albeit limited) influence during the online processing of voice onset time (VOT). Our findings suggest that different kinds of higher-level pragmatic inferences are not uniform in how (successfully) they are integrated with low-level phonetic properties in real time comprehension.

Keywords: pragmatics, contrastive inference, speech perception, VOT, cue integration, eye-tracking

1 Introduction

Listeners have been shown to integrate a vast array of information during speech perception, ranging from low-level acoustic properties of the speech signal, to lexical and morphosyntactic properties of words and higher-level semantic and pragmatic inferences about speaker meaning. Specifically,

lexical status (Ganong, 1980; Samuel, 1981), syntactic category (Isenberg et al., 1980), and semantic congruence (Miller et al., 1984) have all been shown to have an impact on word recognition. Recent work has also probed the interaction of phonetic and pragmatic cues, where pragmatic cues refer to "what is meant beyond what is said" (Grice, 1975). Rohde and Ettlinger (2012) found that pragmatic inferences about coreference have an effect on the perception of a *he-she* continuum. Specifically, the study used implicit causality verbs, which are known to influence expectations about who will be mentioned next in the discourse. Implicit causality verbs introduce strong coreference biases that favor either the subject (*John annoys Tom because he*$_{John}$ ...) or the object (*John annoys Tom because he*$_{Tom}$ ...) interpretation of an ambiguous pronoun (*he*), and listeners have been shown to exhibit such biases in e.g. sentence-completion tasks (Garvey & Caramazza, 1974). Rohde and Ettlinger (2012) rely on the existence of such subject-biasing and object-biasing verbs, and construct *he*-biasing (*Tyler deceived Sue because* □ *couldn't handle a conversation about adultery.*, *Joyce helped Steve because* □ *was working on the same project.*) and *she*-biasing (*Abigail annoyed Bruce because* □ *was in a bad mood.*, *Luis reproached Heidi because* □ *was getting grouchy.*) sentences. Their results indicate that when listening to ambiguous words (*he/she*, marked by □), listeners are indeed more likely to indicate that they heard *he* in *he*-biasing and *she* in *she*-biasing contexts. This suggests, then, that listeners integrate bottom-up phonetic information with high-level (pragmatic) inferences about events, participants, and coreference in discourse.

While the above studies employ offline categorization tasks, the time-course of how listeners disambiguate ambiguous phonetic input based on higher-level information has also been investigated. To begin with, various studies have found that the activation of target words and their competitors shows gradient sensitivity to small, within-category differences in VOT (McMurray et al., 2008, 2002), and that this influence of graded acoustic information can persist over multiple syllables or words (McMurray et al., 2009). McMurray et al. (2009), for example, tested pairs of words such as *parakeet* and *barricade*, which differ in voicing in the initial plosives, but overlap for the subsequent four phonemes. These word pairs created a potential lexical garden-path before the point of disambiguation (i.e. the vowel of the final syllable). The results indicate that the likelihood of initially fixating on the image of the *parakeet* or the *barricade* was affected by the VOT of the initial consonant. Crucially, the time it took for participants to recover from incorrect interpretations (i.e., switch fixations from the competitor to the target) was also a function of VOT, such that the far-

ther the onset phoneme was from the category boundary, the longer the recovery took. This finding constitutes evidence that a continuous representation of VOT is available at the point of disambiguation, suggesting that not only does word recognition show sensitivity to fine-grained subphonetic detail, but this information can also be preserved throughout the processing system. Brown-Schmidt and Toscano (2017) extended McMurray et al.'s (2009) online interpretation and recovery time findings to investigate the integration of graded acoustic information with discourse-level representations. They created a *he-she* continuum and set up the discourse expectation of the pronoun by manipulating whether a referent had been mentioned first or second in a context story. They showed that graded acoustic information exerts a sustained influence. This finding is consistent with the findings of Rohde and Ettlinger (2012), and demonstrates that continuous acoustic differences affect the categorization of pronouns, and this effect is apparent during online processing as well.

The goal of our current study is to further probe the co-influence of acoustics and pragmatics in online comprehension, manipulating a different type of pragmatic information. While previous studies have examined the effect of referential bias driven by either verb bias or discourse salience, we will look at the contrastive function of prenominal adjectives, a case where pragmatic information comes more directly in the form of Gricean reasoning. Using eye-tracking, Sedivy et al. (1999) found that listeners have a robust bias towards interpreting prenominal adjectives contrastively, and that they use the contrastive inference to resolve temporary referential ambiguities in online processing. For example, in one of their experiments (Experiment 1B), a visual display with four objects was presented: two objects were of the same category but differed with respect to a salient property such as color (e.g. yellow comb, pink comb), one object was of a different category but shared a salient property with one member of the minimal pair (e.g. yellow bowl), and there was one unrelated object (e.g. metal knife). Participants were given instructions such as *Touch the yellow comb/bowl*, which were initially (before the noun) compatible with either the yellow comb or the yellow bowl. Sedivy et al. (1999) found that listeners fixated on the target earlier when it was the object with a contrasting pair (i.e. the yellow comb, which had the contrasting pink comb) than when it was the one without a contrasting pair (i.e. the yellow bowl). That is, targets were distinguished from a competitor object faster when a contrasting object was present in the display.

These findings can be given an explanation in terms of Grice's (1975) maxim of quantity. A rational listener assumes that a cooperative speaker

would not produce more information than necessary, and therefore must have had a communicative goal in mind for modifying the object noun. Specifically, when there are two combs in the visual context, the communicative goal is to distinguish the yellow comb from the pink comb. On the other hand, when only one object of a certain category is present in the display (i.e. there is only one bowl), there would have been no communicative advantage in specifying the color of that object. Therefore listeners may reason that a speaker is most likely to use a modifier (e.g. *yellow* in *yellow comb*) when they need to distinguish between two objects from the same category (i.e. yellow comb vs. pink comb) in the visual display. In other words, when faced with a choice between a contrastive and noncontrastive interpretation of adjectival (e.g. color) modification, listeners prefer the contrastive interpretation, and are biased to take the adjective to be referring to the object on the display that has a contrasting pair (see also Sedivy, 2003; Sedivy, 2005; and Aparicio et al., 2015). In our experiment, we capitalize on this bias to interpret adjectives contrastively, and investigate its interaction with the acoustic properties of speech, focusing on the VOT continuum.

2 Experiment: Pragmatic manipulation

In the following we describe a Visual World paradigm eye-tracking experiment (Tanenhaus et al., 1995), in which we investigate the interaction of phonetic cues with Gricean pragmatic reasoning. Our experimental paradigm is inspired by McMurray et al. (2009) and Brown-Schmidt and Toscano (2017) on the acoustic, and by Sedivy et al. (1999) on the pragmatic manipulation side.

If the pragmatic inferences driven by Gricean reasoning about prenominal modification can be recruited immediately to guide speech perception in real-time comprehension, we predict that our pragmatic manipulation would influence how participants perceive ambiguous phonetic input along a VOT continuum. Such an effect might show up in participants' word recognition (categorical behavioral data), and/or in their online processing (eyegaze data). On the other hand, if Gricean pragmatic reasoning is unlike previously investigated forms of pragmatic information (e.g. the bias carried by implicit causality verbs á la Rohde and Ettlinger (2012) or discourse reference context á la Brown-Schmidt and Toscano (2017)), then our pragmatic manipulation would not strongly modulate the bottom-up processing of the acoustic information.

2.1 Participants

Participants were monolingual native speakers of American English recruited from the University of Chicago community. Participants were compensated with $10 or course credit. No participants reported any history of speech or hearing impairments. Data was collected from 40 participants. Participants were excluded from analysis if their behavioral data showed no sensitivity to the VOT manipulation (e.g. participants not reaching at least 50% correct for the two continuum endpoints - 9 subjects), or if track loss from their eye-tracking data exceeded 30% (3 subjects). Following exclusions, data from 28 participants is reported here.

2.2 Materials and procedure

The experiment was administered using a Tobii T120 eye-tracker (screen and resolution: 17-inch TFT, 1280 $\times$ 1024 pixels) and E-Prime 2.0. It employed the Visual World eye-tracking paradigm and had a 7 $\times$ 3 design, which was implemented within participants. We introduced a 7-condition phonetic manipulation (VOT steps) and a 3-condition pragmatic manipulation (visual displays), to be detailed in the following.

In the experiment, a fixation cross was presented before each trial. Once the eye-tracker registered that the participant was looking at the fixation cross, a red border surrounding it appeared. Participants could then proceed to the trial with a mouse click. In each trial, participants were presented with a visual display while hearing a sentence with the form *Click on the ADJ NOUN*, where the NOUN was one of the words from two minimal pairs (*bear/pear*, *bees/peas*), and the ADJ was a monosyllabic color adjective (*red*, *gold*, *grey*, *teal*). All sound editing was performed using Praat (Boersma & Weenink, 2017). The duration of the target adjectives was manipulated by cutting out periods (zero crossing to zero crossing) from the middle of the vowel, so that each adjective would have the same length (233ms). The target stimuli were two 7-step VOT continua (*bear* to *pear*, *bees* to *peas*), where the initial labial of the nouns ranged from /b/ to /p/ in (approximately) 7ms increments. When constructing the audio stimuli, aspiration from the voiceless labial (*p*) was added in increments of 7ms to the corresponding voiced labial (*b*). The 7ms window was adjusted slightly when necessary to ensure that each increment would start and end in a zero-crossing. Target adjectives and nouns were added to the same *Click on the* carrier phrase with matching onset timing for both adjectives and nouns: adjective onsets were timed at 833ms and noun onsets at 1200ms.

To ensure that participants perceived two distinct words, a pause was inserted between the offset of the adjective and the onset of the noun.

As for the pragmatic manipulation, the visual display was presented in three conditions. In all conditions, images were positioned in each of the four corners of the screen (top right, etc.), with one image taking up approximately $\frac{1}{9}$th of the screen - see Figure 1 for a schematic version of the display. Image positions were randomized from trial to trial. All displays contained two objects that share the same color (here: *red bear/pear*), and are hence both temporarily compatible with an instruction *Click on the red...* and can serve as potential targets. The Contrast B (CB) and Contrast

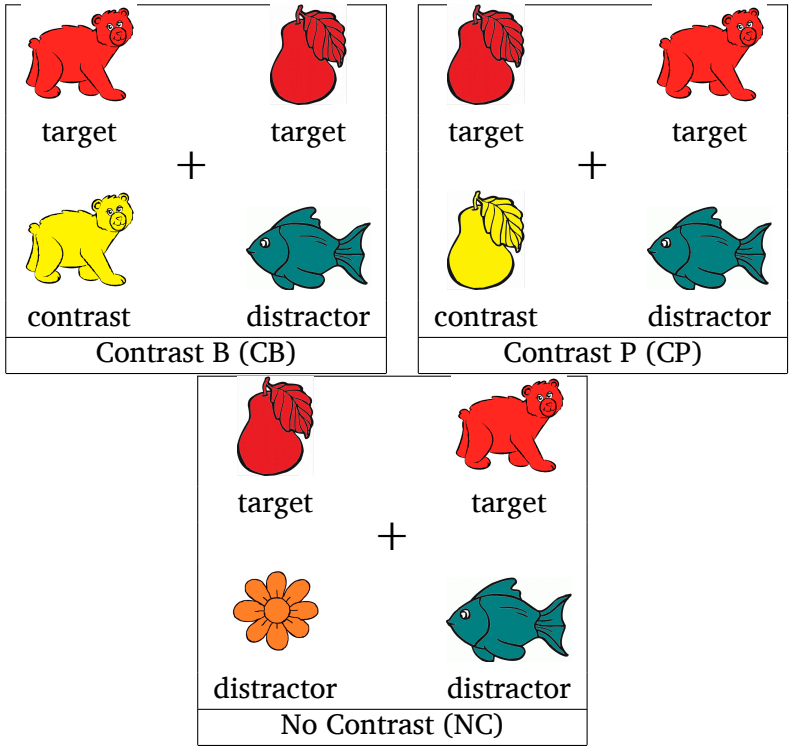

Figure 1: Sample of experimental stimuli. Given the instruction *Click on the red bear/pear*, each condition contains the two potential targets (red bear, red pear). In addition, the CB condition contains a contrast object from the "b" category (yellow bear), and the CP condition from the "p" category (yellow pear). The NC condition contains no contrast object. All conditions contain one/two unrelated distractors

P (CP) conditions contained an additional contrasting object with a different color (here: yellow), where the contrasting object comes either from the "b" category (e.g. bear or bees) or the "p" category (e.g. pear or peas) respectively. The visual displays with contrast objects are expected to trigger pragmatic Gricean reasoning to facilitate the disambiguation of the two potential targets (e.g. Sedivy et al., 1999). Specifically, participants should be biased towards the object that has a contrast comparison. The control condition (No Contrast, NC) contained no contrasting object.

Each participant saw 168 experimental trials (4 color adjectives $\times$ 7 VOT steps $\times$ 3 pragmatic displays $\times$ 2 noun pairs) and 160 filler trials. Filler sentences contained nouns with no bilabial stop onsets (*fox, cup, shoes, house, apple, car, chair, grapes, fish, flower*) and color adjectives different from those used in target sentences (*maroon, yellow, orange, white*). Targets and fillers were produced by a 20-year-old male native speaker in their carrier sentences. Audio stimuli were identical across pragmatic conditions, and had normalized pitch.

2.3 Predictions

Under the hypothesis that pragmatic inferences based on Gricean reasoning can affect the bottom-up phonetic processing, we predict our contrast manipulation to have the following effects. Contrast objects trigger pragmatic Gricean reasoning, which facilitates the disambiguation of two potential targets (note that there is no clear target in the case of ambiguous VOT steps). Therefore, participants should be biased towards the object that has a contrast comparison (i.e. in Figure 1 *bear* under the CB condition and *pear* under CP). This bias may show up in behavioral or online data. As for the behavioral data, we predict different patterns of categorization under the pragmatically biasing conditions as compared to the NC baseline condition, i.e. a larger probability for the same (ambiguous) sound to be categorized as *b* under CB and as *p* under CP relative to the NC baseline. In the eyegaze patterns, we predict the bias to manifest as more looks to the target under CP/CB as compared to NC.

2.4 Results and discussion

Below we analyze and discuss the behavioral click response and the online eye fixation data.

2.4.1 Behavioral data

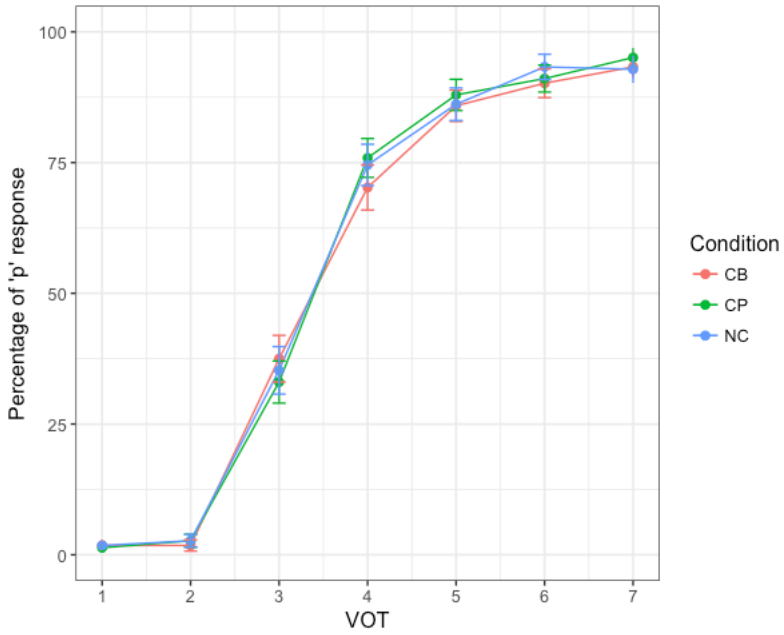

Figure 2: Behavioral results: Summary of percentage of "p" response based on participants' click responses. Error bars show 95% confidence interval.

Figure 2 plots the behavioral click response data: participants' probability of clicking on the "p" objects. A logistic regression model was fitted using the R software, predicting probability of clicking on the "p" objects by VOT step, pragmatic condition, and their interaction. The model only revealed a significant effect of VOT ($p<.0001$), but not the pragmatic contrast manipulation. In other words, participants only showed sensitivity to the acoustic properties of the stimuli, but whether there was a contrasting image had no impact on their categorization behavior (see the identical S-shaped curve for the pragmatic manipulation conditions CP and CB, as well as for the baseline NC).

We also note that p-identification does not reach 100% at the /p/-end of the continuum: even though at step 1 (natural /b/ sound) 0% of participants categorized the sound as "p", at the opposite end, step 7, categorization still does not reach 100%. We believe that this is an artifact of the way our stimuli were constructed. Recall that the original sound was a naturally produced /b/, and aspiration was added to that in 7ms increments to

construct steps 2-7. This means that all steps contain more (e.g. formant) cues for /b/ than for /p/, which has the effect that even the sounds largely perceived as "p" have more suboptimal /p/ acoustic cues than would be expected from a naturally occurring /p/ sound. We return to this asymmetry in the general discussion.

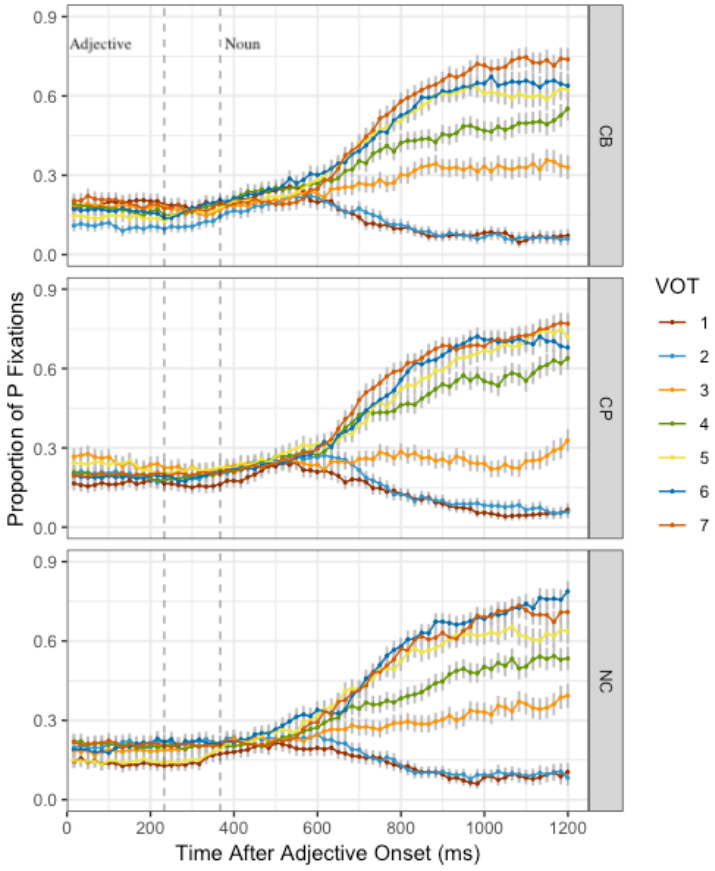

Figure 3: Online results: Proportion of looks to the "p" objects. X axis shows time after adjective onset in ms; vertical lines at 233ms and 367ms represent the offset of the adjective and onset of the noun. Different colored lines represent the 7 VOT steps. Error bars show 95% confidence interval.

2.4.2 Online data

Figure 3 plots the proportion of fixations to the "p" objects as the instruction utterance unfolded in time. Different colored lines represent the different

VOT steps (step 1 is the least /p/-like; step 7 is the most /p/-like), while the three pragmatic conditions are shown on different plots. 0ms marks the onset of the adjective in the acoustic input. The vertical lines mark the offset of the adjective (233ms) and the onset of the noun (367ms) (recall that adjective onsets are timed at 833ms in the total string and noun onsets at 1200ms). All time windows reported below were relative to the actual onset of the adjective in the acoustic input. Since the standard estimate is that it takes 200ms to plan and launch an eye movement (Hallett, 1986), it should be kept in mind that when evaluating timing information to draw conclusions about how quickly our experimental manipulations affect the online processing, the time windows reported below should be adjusted by 200ms. As can be seen in Figure 3, VOT steps start to influence fixations to the "p" objects at a relatively early time point after the adjective onset across all pragmatic conditions, with this effect continuing throughout the trial. The effect of pragmatic condition (comparing the three facets in Figure 3) is more nuanced, but the CP condition does appear to diverge from the NC and CB conditions at a relatively late time window after the adjective onset, at steps 3, 4, 5 (the most ambiguous steps).

To more precisely determine the exact timing of the effect that our pragmatic and phonetic manipulations have on eye movements, we adopted the method introduced by Kingston et al. (2016). For each 100ms time bin from 200ms to 1200ms after the onset of the adjective (excluding the first 200ms because that is the time it takes to plan and launch an eye movement), a logistic regression model was fitted to the proportion of fixations to "p" objects as a function of VOT, pragmatic condition, and their interactions. In addition to the fixed effects, random effects for subjects and items were also included in situations where the models successfully converged. In other words, the same logistic regression model was applied to ten consecutive 100ms time bins in order to track when the effect of our experimental manipulation shows up. Only when an effect consistently appeared for a number of consecutive time windows, were those time windows chosen for follow-up analysis. This method, unlike more arbitrary choices about which time windows to analyze, allows us to track incrementally the effects of phonetics and pragmatics; specifically when they arise and whether they persist.

For each of the 100ms time bin, when a logistic regression model was fit to the data, VOT was set up as a continuous variable in the model, and the three conditions with different visual displays were treatment coded, such that the NC condition was the baseline reference (i.e. the intercept) in the model, and the CB and CP conditions were compared to it. The model

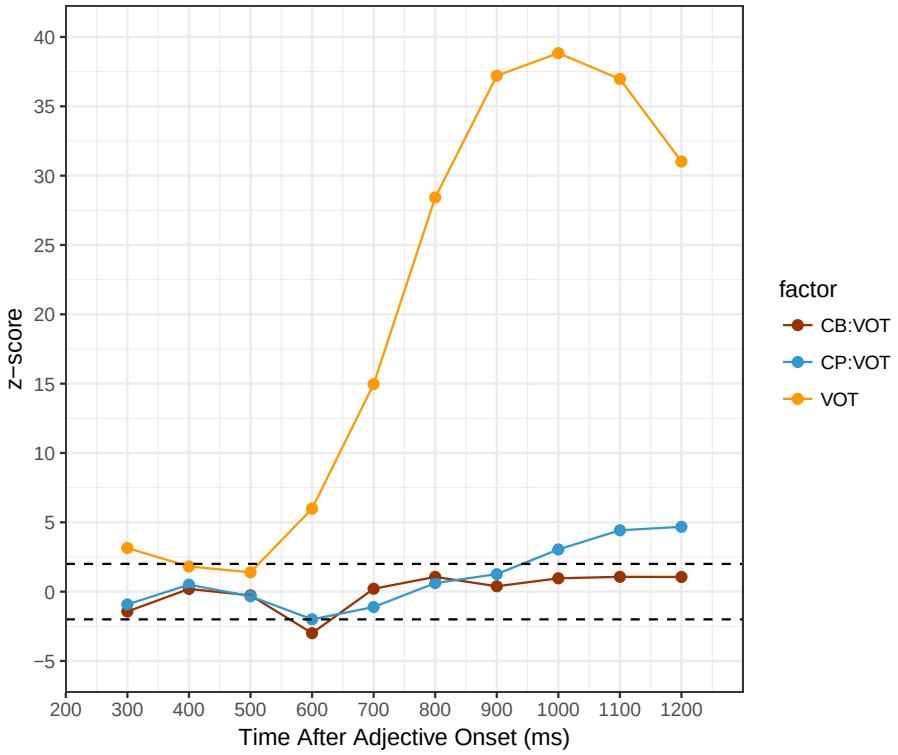

Figure 4: Z-value of coefficients in the model output from models constructed for every 100ms time window. Different colored lines correspond to the main effect of VOT, and its interaction with the two pragmatic conditions (CB:VOT, CP:VOT). The horizontal dashed lines represent $|z|=2$.

output would contain coefficients for a number of different effects (for an example output of such models, see Table 1 below). Since our aim was to investigate whether pragmatic contrast has an effect on speech perception in addition to the acoustic cues (VOT), we were primarily interested in the interaction of the pragmatic conditions (CB/CP) with VOT, and the effect of VOT itself. Figure 4 plots the coefficients from the model outputs, for the critical effects (VOT step, CB:VOT, CP:VOT) in each time bin. Instead of plotting the raw values of these coefficients, we plotted the z-value for each coefficient, with $|z|>2$ as a rough criterion to indicate a significant effect. As expected, the VOT effect starts to become significant in the very early time bin 200-300ms, gets larger during the 500-600 time bin, and continues until the end. The CP:VOT interaction, on the other hand, only starts showing a significant effect at a late time bin (900-1000ms bin), with

Table 1: Parameter estimates, standard errors, z values and p values from a logistic regression model of the proportion of looks to the "p" objects in the 200-1200ms time window

	Estimate	Std. Error	z value	p value
(Intercept)	-1.8786	0.0194	-96.82	$<.001$
CB	-0.0879	0.0278	-3.16	$<.01$
CP	-0.0078	0.0274	-0.28	0.7764
VOT	0.2601	0.0040	64.51	$<.001$
CB:VOT	0.0016	0.0058	0.27	0.7849
CP:VOT	0.0082	0.0057	1.44	0.1493

this effect continuing until the end of the trial. The CB:VOT interaction shows no significant effect.

Motivated by the consecutive time bin analysis showing an early and persisting effect of VOT, we analyzed fixations during the collapsed 200-1200ms time window. This is also in line with the time windows chosen in previous work (i.a. Brown-Schmidt and Toscano, 2017), which analyze the whole trial, taking into account the time it takes to plan eye movements. A logistic regression model was fitted to the eye-tracking data, with fixations to the "p" objects as the response variable - see Table 1. As before, VOT was set up as a continuous variable in the model, and the three conditions with different visual displays were treatment coded, such that the NC condition was the baseline reference (i.e. the intercept) in the model, and the CB and CP conditions were compared to it. Similarly to the click response data, we found a robust effect of VOT ($p<.0001$). There were no reliable effects of pragmatics at earlier time windows: investigating the full 200-1200ms time window, we find that only VOT is a significant predictor.

Given the effect present in Figure 4, we conducted a further analysis for a narrower time window, and also collapsed the window that is 900-1200ms after the adjective onset (which is about 500ms after the onset of the noun) - see Table 2. In this analysis, we found a significant interaction between VOT steps and looks to the "p" objects in the CP vs. NC comparison. In particular (see Figure 5), there were more looks to the "p" objects in the CP condition as compared to the NC condition at VOT steps 4 ($p<.001$), 5 ($p<.001$) and 7 ($p<.05$). The effect is not significant at step 6 ($p=.902$). In other words, for steps 4-7, which were steps predominately categorized as "p" in the click response data, we observe a facilitatory effect of the CP visual display. At VOT steps 1-3, which were predominately categorized as "b", we find an inhibitory effect of CP, i.e. there are less looks to the

"p" objects in CP as compared to NC (steps 1 ($p<.001$), 2 ($p<.05$) and 3 ($p<.001$)). As mentioned, no reliable CB:VOT interaction was observed, meaning that the CB pragmatic manipulation did not have the effect that CP did. We will return to this asymmetry in the general discussion below.

Table 2: Parameter estimates, standard errors, z values and p values from a logistic regression model of the proportion of looks to the "p" objects in the 900-1200ms time window

	Estimate	Std. Error	z value	p value
(Intercept)	–2.5565	.0043	–59.34	$<.001$
CB	–.1854	.0622	–2.98	<0.01
CP	–.3738	.0627	–5.96	$<.001$
VOT	.5562	.0092	60.24	$<.001$
CB:VOT	.0240	.0133	1.81	0.0706
CP:VOT	.0904	.0135	6.68	$<.001$

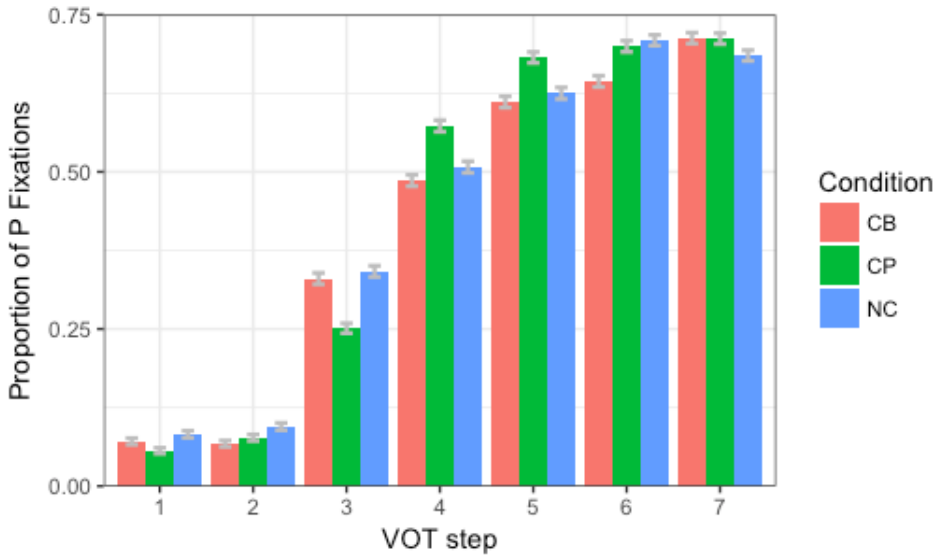

Figure 5: Online results: Proportion of looks to the "p" objects as a function of VOT step in the 900-1200ms time window. Different colors represent the three pragmatic conditions (CB, CP, NC) at each VOT step. Error bars show 95% confidence interval.

The analysis reported above pooled together fixations at each VOT step irrespective of whether a participant's eventual click-choice was biased to-

wards “p” or “b” at that step. Per the suggestion of one reviewer, we also conducted an additional analysis (following McMurray et al., 2002), taking into account participants’ decision bias at different VOT steps. This amounted to dividing the data into two sets: trials in which the target was predominately ($\geq$50%) interpreted as “b” (steps 1-3), and trials in which the target was predominately interpreted as “p” (steps 4-7). For each of the two subsets, we excluded data from any trial on which the participant clicked on the “competitor” picture (i.e. the one representing the opposite side of the category boundary: a “p” picture for steps 1-3 or a “b” picture for steps 4-7). Because in the analysis presented in Figure 4, we saw that pragmatic effects only interacted with VOT in a late time window, we restrict our analysis here to the same late time window: 900-1200ms. Logistic regression models were fitted, predicting fixations to the target object (“b” for steps 1-3 and “p” for steps 4-7) by VOT, pragmatic condition, and their interaction. This analysis revealed that the effect of pragmatics in the CP condition (i.e., CP:VOT interaction) is significant ($p<.001$) for both steps 1-3 and 4-7. We also found a more limited effect of CB (i.e. CB:VOT interaction, $p<.01$) for steps 4-7 (“p” target) only. VOT itself also remains

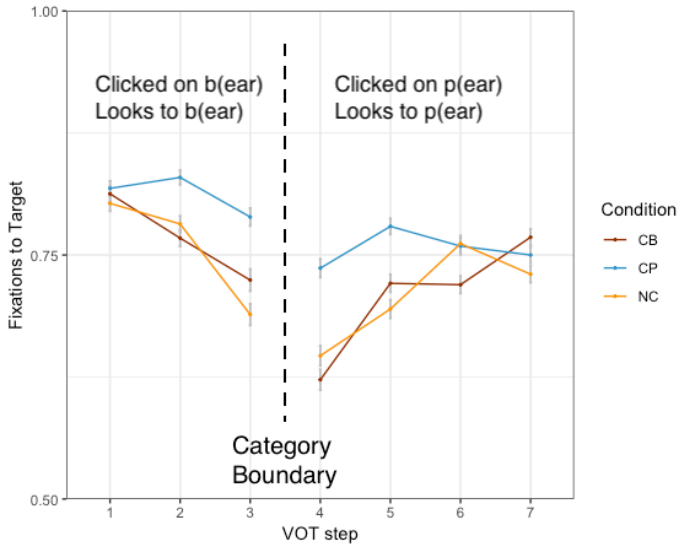

Figure 6: Online results: Proportion of looks to the target objects (“b” at steps 1-3 and “p” at steps 4-7) as a function of VOT step in the 900-1200ms time window. Different colors represent the three pragmatic conditions (CB, CP, NC) at each VOT step. Error bars show 95% confidence interval.

a significant predictor for both steps 1-3 and 4-7 on this analysis as well ($p<.001$).

Overall, separating "b-clicked" from "p-clicked" trials revealed an effect largely consistent with our initial analysis. We observed an effect of pragmatics in the late 900-1200ms time window, but this effect is still largely asymmetric, and more evident in the CP than in the CB pragmatic condition. Similarly to the initial analysis, the effects of pragmatics mainly manifested as more looks to the "p" objects in CP as compared to NC at the ambiguous but yet "p-biasing" VOT steps 4 ($p<.001$) and 5 ($p<.001$). Also similar to the initial analysis, there were more looks to the "b" objects (i.e. fewer looks to the "p" object) in the CP condition as compared to the NC condition at VOT step 1 ($p<.05$), 2 ($p<.001$) and 3 ($p<.001$). Although the new analysis revealed an overall effect of CB:VOT interaction for steps 4-7, we would caution against over-interpreting this in the absence of any follow-up studies. As revealed by Figure 6, there is not a clear consistent influence of CB across different VOT steps. We therefore will not discuss the effect of CB further.

3 General discussion

While Rohde and Ettlinger (2012) capitalized on the pragmatic information carried by implicit causality verbs, and found that it affects inferences about coreference in the face of ambiguous acoustic stimuli, our study differs from it in the kind of pragmatic knowledge that is tapped into. Specifically, we were interested in whether pragmatic reasoning based on Grice's (1975) maxim of quantity, as manifested in the interpretation of prenominal modification, could guide the perception of acoustic information. Our offline behavioral judgment results found no direct effect of this specific kind of pragmatic inference on phonetic categorization, which is instead completely determined by the acoustic cues (VOT, Figure 2). This is in contrast with previous results showing that phonetic categorization could be influenced by information about lexical status, syntactic category, semantic congruity and coreference relations in a discourse context. Although the behavioral categorization judgment did not reveal any effect of Gricean pragmatic reasoning, the online eye-tracking data showed some evidence for pragmatic influence. However, the effect of pragmatics (i.e. a bias from the contrasting object when it comes to the interpretation of ambiguous sounds) is constrained in that it appears much later than the effect of phonetic cues. Altogether, both the behavioral and online results suggest only

a limited influence of (Gricean) pragmatics.

One explanation for the difference between the current study and previous studies – which found a more robust interaction between phonetic and higher-level linguistic cues – is that the nature of the pragmatic information we examined is in some way fundamentally different from previously investigated cues. In particular, the pragmatic information that Rohde and Ettlinger (2012) manipulated (i.e., the referential bias introduced by implicit causality verbs) ultimately derives from lexical items, that is, the verbs themselves. Similarly, information from lower levels of the linguistic structure, such as lexical status or syntactic category, is also information encoded in particular lexical items. On the contrary, the contrastive inference of prenominal adjectives is carried neither by the adjective, nor by the noun. Rather, it arises from Gricean pragmatic principles. Rational hearers assume that their interlocutors are cooperative and have some communicative goal in mind, and would therefore only use prenominal modification if it served the purpose of identification. This reasoning process leads hearers to preferentially assign a contrastive interpretation to adjectives. One possible way to reconcile the findings of the current study with previous work is to argue that the pragmatic cues triggered by local information encoded in lexical items might exert a stronger influence on gradient phonetic categorization, as compared to pragmatic cues that are derived from global computation about the communicative goals. Nevertheless, further research is needed, particularly on the integration of a wider range of contextual/pragmatic cues in phonetic categorization, in order to ascertain the validity of this hypothesis.

The albeit limited influence of pragmatic context that we observed nonetheless imposes constraints on the overall architecture of language processing. The effect of pragmatic inferences (evidenced by proportion of looks to the target) arrives at a very late stage in processing, after the participant has heard the noun in the instructions. This is despite the fact that in our experimental paradigm, pragmatic information arrives early, since it is carried by the prenominal adjective. Thus, for any language processing model to be compatible with the current findings, it needs to implement interactions between different linguistic levels, but at the same time assign differential strength to these interactions. Models with these features have been proposed before: Kawamoto (1988, 1993) developed a model in which associations between different types of lexical information (orthography, phonology, alternative meanings) are stronger than the association between contextual information and any individual meaning of the lexical item. This allowed for the computation of word meaning to be influenced

by contextual information, but crucially, multiple meanings were still computed even when there was contextual bias favoring only one of them (see also *bottom-up priority* in the sense of Marslen-Wilson, 1987). Similarly, at the sentence parsing level, MacDonald et al. (1994) also developed a model that is contextually constrained, but lexically dominated. Models of this kind allow room to handle different varieties of contextual effect. Future research needs to examine whether these models can derive the current finding that contextual cues arising from Gricean pragmatic reasoning (unlike pragmatic cues ultimately rooted in lexical properties) are slow to be integrated with bottom-up acoustic cues.

Additional to our main finding that the influence of Gricean reasoning on speech perception is limited, there remain some empirical puzzles regarding the precise nature of the observed pragmatic effect. First, there is a persisting asymmetry between the two ends of the VOT contrast: pragmatic effects are largely constrained to the CP display. That is, the perception of the voiceless "p", but not the voiced "b", is affected by top-down pragmatic influence. We argue that the reason for this is that "p" had more incongruous acoustic cues (since in our stimuli "p" tokens were modified from the "b" tokens), hence there was more room for pragmatics. That is, pragmatics only introduces a bias in participants' speech perception when phonetic information is suboptimal. Second, the effect of pragmatics is conditioned on the behavioral response. At VOT steps 4-7, we found more looks to the "p"-target object in the CP, as compared to the NC baseline condition. At these steps, phonetic and pragmatic cues point in the same direction, and thus facilitation is predicted. Sounds that are perceived as "p" in the behavioral data, coupled with a "p"-biasing pragmatic information led to faster target identification. At steps 1-3, there were more looks to the "b"-target object in the CP condition, relative to the NC baseline. At these steps, phonetic and pragmatic cues are in conflict: pragmatic information supports a "p"-interpretation, but the initial plosive is eventually categorized as "b". This incongruence may have led to uncertainty regarding target identification, resulting in more looks (relative to the NC baseline condition) to the competitor object that is supported by the acoustic information. These tentative proposals would require verification in future follow-up studies.

4 Conclusion

Existing work has shown that listeners integrate cues from multiple levels of linguistic structure. Specifically, these levels and information sources

are known to range from phonetic, through lexical and morphosyntactic, to semantic. Recently, Rohde and Ettlinger (2012) have shown that even pragmatic inferences about upcoming coreference (relying on the properties of subject- vs. object-biasing implicit causality verbs) can have an impact on speech perception. This constitutes evidence that the maximum range of linguistic cue integration involves phonetics and pragmatics, the two most disparate domains of linguistic structure.

In our study, we probed further whether different kinds of pragmatic information can uniformly be integrated with bottom-up acoustic information. We capitalized on listeners' known preference to interpret prenominal adjectives contrastively (as evidenced by eye fixations), and investigated whether this preference has an effect on the perception of a /p/-/b/ continuum. We found no effect of the pragmatic manipulation on behavioral judgments on phonetic categorization, and that pragmatics had a limited influence during online processing. Our results suggest that pragmatic cues about the contrastive function of adjectival modification are secondary to the bottom-up acoustic information during speech perception. This finding informs our understanding of the limits of linguistic cue integration: it provides evidence that not all kinds of pragmatic information can exert an immediate top-down effect on speech perception.

References

Aparicio, H., Xiang, M., & Kennedy, C. (2015). Processing gradable adjectives in context: A visual world study. In S. D'Antonio, M. Moroney, & C.-R. Little (Eds.), *Semantics and linguistic theory (SALT)* (Vol. 25, pp. 413–432). LSA and CLC Publications. doi: 10.3765/salt.v25i0.3128

Boersma, P., & Weenink, D. (2017). *Praat: doing phonetics by computer [computer program]*. `http://www.praat.org/`.

Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, *32*(10), 1211–1228. doi: 10.1080/23273798.2017.1325508

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125. doi: 10.1037/0096-1523.6.1.110

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, *5*(3), 459–464.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, volume 3: Speech acts* (p. 41–-58). New York: NY: Academic Press.

Hallett, P. E. (1986). Eye movements. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (p. 10.1–-10.112). New York: Wiley.

Isenberg, D., Walker, E. C. T., & Ryder, J. M. (1980). A top-down effect on the identification of function words. *The Journal of the Acoustical Society of America*, *68*(S1), S48–S48. doi: 10.1121/1.2004759

Kawamoto, A. H. (1988). Interactive processes in the resolution of lexical ambiguity. In S. I. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical ambiguity resolution: Computational linguistic, and psychological perspectives* (pp. 195–228). New York: Morgan Kaufmann.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, *32*(4), 474–516.

Kingston, J., Levy, J., Rysling, A., & Staub, A. (2016). Eye movement evidence for an immediate Ganong effect. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(12), 1969–1988. doi: 10.1037/xhp0000269

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. doi: 10.1037/0033-295X.101.4.676

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102. (Special Issue Spoken Word Recognition) doi: 10.1016/0010-0277(87)90005-9

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1609–1631. doi: 10.1037/a0011747

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42. doi: 10.1016/S0010-0277(02)00157-9

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category vot affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91. doi: 10.1016/j.jml.2008.07.002

Miller, J. L., Green, K., & Schermer, T. M. (1984). A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception & Psychophysics*, *36*(4), 329–337.

doi: 10.3758/BF03202785

Rohde, H., & Ettlinger, M. (2012). Integration of pragmatic and phonetic cues in spoken word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(4), 967-–983. doi: 10.1037/a0026786

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychcology: General*, *110*(4), 474–494. doi: 10.1037/0096-3445.110.4.474

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23. doi: 10.1023/A:1021928914454

Sedivy, J. C. (2005). Evaluating explanations for referential context effects: Evidence for gricean mechanisms in online language interpretation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 345–364). Cambridge, MA: The MIT Press.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147. doi: 10.1016/S0010-0277(99)00025-6

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. doi: 10.1126/science.7777863