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

Eszter Ronai¹,*, Yenan Sun¹, Alan C. L. Yu¹, and Ming Xiang¹

¹Department 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 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\textit{John} ...) or the object (John annoys Tom because \textit{he\textsubscript{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.

To investigate whether the previously observed effect is a general finding about how pragmatic information modulates bottom-up speech perception, or whether the effect is constrained to a particular kind of pragmatic cue (i.e. the referential bias driven by the verb bias), the current study manipulates a different type of pragmatic cue. Specifically, we 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, Experiment 1B) found a robust bias in the way listeners interpret adjectivally modified nouns contrastively and that the contrastive information is used to resolve temporary referential ambiguities in online processing. In their study, participants saw a display with four objects: 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 heard instructions such as
Touch the yellow comb or Touch the yellow bowl. When interpreting prenominal adjectives, listeners might assume, following Grice’s (1975) maxim of quantity, that the speaker would not be giving them more information than necessary. Therefore listeners may reason that the modifier (e.g. yellow in yellow comb) is needed to pick out the relevant item, which leads them to expect that there are probably other such items (i.e. another comb) in the visual display. Thus, a listener who hears instructions involving the contrasting object (yellow comb) would have faster looks to the target than when she hears instructions involving the competitor object (yellow bowl). The instruction Touch the yellow ... is temporarily compatible with either the yellow comb or the yellow bowl, but because only the yellow comb has a contrasting pair (the pink comb), listeners fixate on that target earlier. That is, targets are distinguished from a competitor object faster when a contrasting object is present in the display.

As mentioned, these findings can be given a Gricean explanation: the listener assumes that the speaker must have had a communicative goal in mind for modifying the object noun, and this goal may have been 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, 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 for an extension of the paradigm to relative and absolute adjectives). 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, in which we investigated the interaction of phonetic cues with Gricean pragmatic reasoning.
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. Participants were excluded from analysis if their behavioral data showed no sensitivity to the VOT manipulation (that is, they perceived all sounds on the continuum as \( b \) or \( p \)), or if track loss from their eye-tracking data exceeded 30%. Following exclusions, data from 28 participants is reported here.

2.2 Materials and procedure

The experiment used 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, participants were presented with a visual display while hearing a sentence with the form \( \text{Click on the ADJ NOUN} \), where the NOUN was one of the words from two minimal pairs (\( \text{bear/pear, bees/peas} \)), and the ADJ was a monosyllabic color adjective (\( \text{red, gold, grey, teal} \)). The duration of the target adjectives was manipulated so that each adjective would have the same length (233ms). The target stimuli were two 7-step VOT continua (\( \text{bear to pear, bees to peas} \)), where the initial labial of the nouns ranged from /b/ to /p/ in 7ms increments. When constructing the audio stimuli, aspiration from the voiceless labial (\( p \)) was added in increments of 7ms to the corresponding voiced labial (\( b \)). Target adjectives and nouns were added to the same \( \text{Click on the} \) carrier phrase with matching onset timing for both adjectives and nouns (adjective onsets were timed at 833ms and noun onsets at 1200ms). Filler sentences contained nouns with no bilabial stop onsets (\( \text{fox, cup, shoes, house, apple, car, chair, grapes, fish, flower} \)) and disyllabic color adjectives (\( \text{maroon, yellow, orange, white} \)). 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.

As for the pragmatic manipulation, the visual display (Figure 1) was presented in three conditions. All of them contained two objects that share the same color (here: \( \text{red bear/pear} \)), and are hence both temporarily compatible with an instruction \( \text{Click on the red ...} \) and can serve as potential targets. The Contrast B (CB) and Contrast 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.

![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.](image)
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. bear/bees under the CB condition and pear/peas under CP). This bias may show up in behavioral or online data. As for the behavioral data, we predict different categorizations 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 plots the behavioral click response data: participants’ probability of clicking on the “p” objects. A logistic regression model was fit using the R software, predicting probability of clicking on the “p” objects by VOT step and pragmatic condition. 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 /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
Figure 2: Behavioral results: Summary of percentage of “p” response based on participants’ click responses. Error bars show 95% confidence interval.

/p/, which has the effect that even the sounds largely perceived as “p” have more conflicting and suboptimal /p/ acoustic cues than would be expected from a naturally occurring /p/ sound. We return to this observation in our analysis of the eye fixation data.

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 red vertical line marks the offset of the adjective, and noun onsets are at 367ms (recall that adjective onsets are timed at 833ms in the total string and noun onsets at 1200ms). All the 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 150-200ms to plan and launch an eye movement, it should be kept in mind that when evaluating timing information to draw conclusions about
Figure 3: Online results: Proportion of looks to the “p” objects. X axis shows time after adjective onset in ms, with the red vertical dashed line representing the offset of the adjective. Different colored lines represent the 7 VOT steps. Error bars show 95% confidence interval.

how quickly our experimental manipulations affect the online processing, the time windows reported below should be offset by 200ms. A logistic regression model was fit to the eye-tracking data, with fixations to the “p” objects as the response variable. 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 find a robust effect of VOT across all pragmatic conditions, starting from 500ms after the adjective
In order to pinpoint the time window where there is an additional effect of pragmatics on low level speech perception, we fit logistic regression models on 10 consecutive 100ms time windows (from 200ms to 1200ms). These models predicted 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. Since our aim was to investigate whether pragmatic contrast has an effect on speech perception in addition to the acoustic cues (VOT), we were interested specifically in the interaction of the pragmatic conditions (CB/CP) with VOT. Figure 4 plots the coefficients from the model outputs, for the critical effects (VOT step,
CB:VOT, CP:VOT) in each time bin. For each coefficient, we take $|z| > 2$ as a rough criterion to indicate a significant effect (Kingston et al., 2016). As expected, the VOT effect becomes significant in the very early time bin, 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, starting around 900ms to the end. The CB:VOT interaction shows no significant effect.

To reiterate, there were no reliable effects of pragmatics at earlier time windows. However, given the effect present in Figure 4, we also collapsed the time window that is 900-1200ms after the adjective onset (which is about 500ms after the onset of the noun). In this analysis, we find a significant interaction between VOT steps and looks to the “p” objects in the CP vs. NC comparison (Table 1). 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 on step 6 (p = .902). In other words, because steps 4-7 are the ambiguous sounds categorized as “p” anyway in the click response data, we observe a facilitatory effect of CP at them. At VOT steps 1 (p < .001), 2 (p < .05) and 3 (p < .001), which are 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. As mentioned, no reliable CB:VOT interaction was observed, meaning that the CB pragmatic manipulation did not have the effect that CP did. We take this difference between “p” and “b” to be an artifact of our stimuli construction, similarly to p-identification not reaching 100% in the click response data. In particular, sounds perceived as “p” have suboptimal /p/ acoustic cues, but this is not the case for “b” sounds - a point we return to in the general discussion below.

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 900-1200ms time window

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.5565</td>
<td>.0043</td>
<td>-59.34</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>CB</td>
<td>-1.854</td>
<td>.0622</td>
<td>-2.98</td>
<td>.0028</td>
</tr>
<tr>
<td>CP</td>
<td>-3.738</td>
<td>.0627</td>
<td>-5.96</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>VOT</td>
<td>.5562</td>
<td>.0092</td>
<td>60.24</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>CB:VOT</td>
<td>.0240</td>
<td>.0133</td>
<td>1.81</td>
<td>0.0706</td>
</tr>
<tr>
<td>CP:VOT</td>
<td>.0904</td>
<td>.0135</td>
<td>6.68</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>
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.

3 General discussion

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 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 asked the question of whether different kinds of pragmatic information can uniformly be integrated with bottom-up acoustic information. Specifically, we capitalize on listeners’ known preference to interpret prenominal adjectives contrastively (as evidenced by eye fixations), and investigate whether this preference has an effect on the perception of a /p/-/b/ continuum.

Our results found no direct effect of pragmatic contrast on phonetic categorization output (behavioral judgment data), which is instead completely
determined by the acoustic cues (VOT). The eye-tracking data showed some evidence for pragmatic influence on consonant categorization, but the effect is highly constrained. First, the effect of pragmatics (i.e. a bias from the contrasting object when it comes to the interpretation of ambiguous sounds) appears much later than the effect of phonetic cues. Second, there is an asymmetry between the two ends of the VOT contrast: the perception of the voiceless “p”, but not the voiced “b”, is affected by top-down pragmatic influence (only the CP:VOT interaction is significant). We argue that the reason for this is that “p” had more incongruous acoustic cues (since “p” tokens were modified from the “b” tokens), hence there was more room for pragmatics. The precise nature of the pragmatic effect of CP is conditioned on the behavioral response, such that we observe facilitation when phonetic and pragmatic cues point in the same direction (i.e. sounds that are perceived as “p” in the behavioral data). Yet there is inhibition when the two cues are incongruous (sounds perceived as “b”).

4 Conclusion

We know from earlier studies that different stages of linguistic processing (from phonetic to pragmatic) are not completely autonomous. Such top-down effects are compatible with either an interactive model of cognition (where top-down and bottom-up factors interact directly, e.g. TRACE, McClelland and Elman, 1986), or a modular model (where top-down factors only have an impact post-perceptually, e.g. MERGE, Norris et al., 2000). In this paper, however, our primary goal was not to weigh in on issues of architecture, rather to probe further the question of the maximal range of information sources that listener can integrate, and to look at the effect of a previously uninvestigated kind of pragmatic information on spoken word recognition.

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 be integrated with phonetic information.

We observe no effect of pragmatics on behavioral judgments, even though previous results about the effects of lexical status, syntactic category, semantic congruity and pragmatic inferences about coreference all showed
up in such measures. We observe a limited effect of pragmatics on eye fixations, but only in cases when the acoustic information contained suboptimal cues. Altogether, 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


