Homophone Disambiguation and Vocal Stereotypes

Julie E Boland

University of Michigan

Shana'e N. Clark

Western Michigan University

Address Correspondence to:
Julie E Boland
Dept of Psychology
University of Michigan
Ann Arbor, MI 48109
jeboland@umich.edu

Abstract

Three visual world eye-tracking experiments investigated how listeners use social cues conveyed by the speaker's voice to recognize the intended meaning of a homophone. For example, in Experiment 2, we measured the time to fixate a (female-biased) fingernails image when "Look at the nails" was spoken in either a woman's voice or man's voice. The image appeared, along with a phonological competitor and two unrelated images, at the onset of the homophone (*nails*). In Experiment 1, the carrier phrase was a biasing sentence context (e.g. "At the end of the month, I write out a check" in either an adult voice or a child voice). In Experiments 2 and 3, we used a "Look at the…" context with and without an overlay of crowd noise, respectively. In Experiment 1, the target image was fixated more quickly when the voice was consistent with the social bias of the homophone image, whereas no such speaker congruity effects were observed in Experiments 2 and 3. These results support a processing account in which social stereotypes associated with the speaker's voice most strongly influence homophone resolution in contexts that encourage predictive processing. We found no support for a bottom-up account in which congruent speaker tokens of a homophone are recognized more quickly because they are a better acoustic match to salient stored exemplars.

Keywords: lexical ambiguity, homophone, speaker effects, eye-tracking, exemplar models, predictive processing

Acknowledgments: We thank Kathryn Couger, Chelsea Prush, Maryam Seifeldin, Jessica Silfen, and Danielle Flanders for help in running the experiments and normative studies. We thank Nick Ellis, for suggesting Experiment 3, and the U of M Psycholinguistics group for feedback on the project.

Homophone Disambiguation and Vocal Stereotypes

Sentence context is an important cue for identifying the meaning of a lexically ambiguous word. For example, in (1), the sentence context most strongly supports the fingernail meaning of "nails." Most scholars agree that biasing sentence context rapidly guides lexical ambiguity resolution, although the relative frequency of the alternative meanings also clearly plays a role, and there is some debate as to how context and meaning frequency interact (e.g., Chen & Boland, 2008; Leinenger & Rayner, 2013; Martin et al., 1999; Sereno et al., 2003; Vu et al., 2003). In this paper, we consider another type of context, the voice of the speaker. A speaker's voice can convey (or index) both individual identity and some of the social categories to which the speaker belongs (e.g., gender, age group, social class, geographic region). For homophones with stereotypical biases, the social identity of the speaker is potentially relevant to resolving the ambiguity. In our study, we investigated whether the speaker's gender and age group influence the speed of lexical ambiguity resolution in spoken sentences. For example, will listeners recover the intended meaning of "nails" in (1) more rapidly when spoken in a woman's voice compared with a man's voice, given that the fingernails meaning is more strongly associated with women? And if listeners do use talker-specific social cues, do such cues have more impact in a linguistically rich sentence like (1) or a neutral carrier sentence like (2)?

1) *I will grab the splinter with my nails.*
2) *Look at the nails.*

Prior research has investigated talker-specific voice effects, both at the level of speech perception and spoken word recognition, and at the level of semantic processing of spoken sentences. We will briefly review that literature before turning to our experiments.

Speech Perception & Spoken Word Recognition. There is now a large body of research investigating how the talker information that is carried in the speech signal is used for recovering linguistic content during speech perception and spoken word recognition. Traditional views assumed that speaker normalization processes filtered out talker-specific aspects of the acoustic signal in order to recognize phonemes and words (e.g., Goldinger et al., 1996; Klatt, 1989; Mullenix & Pisoni, 1990; Ladefoged & Broadbent, 1957). In contrast, exemplar models in which talker information provides cues used to recover the linguistic content are now common (e.g. Johnson 1997; Goldinger, 1998; Hay & Walker, 2011).

Evidence supporting exemplar models includes the finding that listeners are better at processing stimuli spoken by a constant talker rather than multiple talkers (e.g. Creel & Tumlin, 2011; Creel et al., 2008; Nygaard & Pisoni, 1998; Kraljic & Samuel, 2007; Palmeri, Goldinger, & Pisoni, 1993), as well as findings that social categories indexed by the voice (e.g., gender, region) influence how listeners categorize phonemes (Strand & Johnson, 1996) and the speed with which they make lexical decisions (Hay & Walker, 2011). For example, Hay and Walker found that listeners were faster and more accurate to discriminate words from nonwords when the word's age (old or young, estimated from spoken corpora collected at different times) matched the speaker's age (old = 50 years, young = 22 years). This finding supports an exemplar model of the lexicon, in which lexical entries are comprised of acoustically rich exemplar tokens, representing prior encounters with the word. If we have experienced older words as having been uttered primarily by older speakers, a new token of an old word will be more easily recognized when spoken in a voice that is acoustically similar, i.e., old-sounding.

Sumner et al. (2014) also predict speaker effects under some circumstances, but they postulate a dual route theory in which acoustic patterns are mapped to linguistic representations

along two routes, one of which makes use of social features. This allows the theory to predict that listeners will recognize typical and atypical phonetic variants of a word equally quickly, yet remember the less frequent, socially idealized pronunciation better in long term memory tasks. For example, listeners recognize “splinter” equally quickly when pronounced with and without the “t” sound, but they remember the pronunciation with the “t” better. In contrast to exemplar theories, linguistic encoding and socio-acoustic encoding occur in separate streams of processing, and thus are dissociable, though they are expected to interact. For example, strongly female-associated words should be recognized more quickly in a female voice than a male voice due to the “social weighting” of salient tokens that occurs via the socio-acoustic route. Another way in which this account differs from exemplar theories is that social effects are predicted to be independent of gendered-usage frequency counts. Sumner et al. warn that socio-acoustic effects may be masked under some circumstances, although it is not clear why. They suggest that socio-acoustic effects on word recognition will be most robust “in longer utterances, at the ends of experiments, or in words that slow linguistic processing (like words with late disambiguation points)” (p. 8).

An important question concerns how early in processing talker-specific vocal cues become available. For example, McLennan and Luce (2005) found no talker-specificity effects when listeners performed an easy lexical decision task, but did they find talker-specificity effects when discriminating words from nonwords was difficult, and correspondingly, response times were 30-40 ms slower. Lexical decision times in the primed conditions averaged 759 ms with easy nonwords and 790 ms with difficult nonwords; in the control condition, mean response times were 800 and 837 ms respectively. McLennan and Luce replicated this finding with a shadowing task in which participants had to repeat a spoken word as quickly as possible (no

talker-specificity effects) or in response to a prompt 150ms after word offset (talker-specificity effects observed. They concluded that talker specificity effects are slow to emerge, compared with linguistic effects, during the perceptual processing of spoken words. We will consider the analogous timing issue for the talker-specific social category effects investigated here when considering the modulation of speaker congruity effects we observed across experiments

Semantic Processing. At the sentence level, van Berkum et al. (2008) found that the social category of the speaker, as indexed by voice, determined whether a semantic anomaly effect (the N400) was observed when measuring event-related brain responses. An N400 effect is typically observed about 400ms after a semantically anomalous word compared with a more predictable word, as in (3). Van Berkum et al. observed an effect with the same time-course and scalp distribution, but smaller amplitude, when the critical word was inconsistent with the speaker's social identity. For example, the speaker incongruity N400 effect was observed at "wine" in sentences like (4), spoken in a child's voice, compared with the same sentence spoken in an adult's voice. Crucially, both the semantic anomaly effects and the speaker incongruity effects began to emerge very rapidly, within 200-300 ms after the onset of the critical word. The speaker incongruity effect has also been found with fMRI (Tesink et al., 2009).

*3) You wash your hands with **horse/soap** and water.*

4) *Every evening I drink some **wine** before I go to sleep.*

The time-course of van Berkum et al.'s (2008) speaker incongruity effects suggests that social properties of the speaker's voice may be available in time to influence lexical ambiguity resolution. However, the mechanism through which the speaker's voice is integrated into the message-level representation of the sentence is unclear. The mechanism could be largely predictive, with listeners anticipating what a particular speaker is likely to say. This explanation

is consistent with prior research showing a strong correlation between predictability of a word (usually measured via its cloze value) and the amplitude of the N400 for that word. The predictive mechanism is also consistent with the difference in amplitude between the semantic anomaly sentences (3) and the speaker incongruity sentences (4). For sentences like (3), there was a large N400 to *horse*, but hardly any negative-going activity to the highly predictable "soap". In contrast for sentence like (4), there was a substantial N400 to "wine" for both congruent and incongruent voices, suggesting that it was not highly predictable for either voice, though the N400 was significantly larger for the incongruent voice. If the primary mechanism for the speaker incongruity N400 is predictive processing, it may depend upon the use of self-referential pronouns, which were used in many of van Berkum et al.'s stimuli, such as (4).

Alternatively, the speaker incongruity N400 found by van Berkum et al. (2008) may reflect difficulty at lower levels of processing. As suggested by Hay and Walker (2011), it may be more difficult to access the word "wine" when spoken in a child's voice, if one's prior experience with the word has been primarily in adult voices. Such a mechanism does not rely upon any high-level predictive processing; it is simply a side effect of bottom-up word recognition under an exemplar-based theory of the lexicon. In turn, any difficulties in lexical access may interfere with semantic integration of the word into the sentence and/or discourse model. Of course, these two mechanisms are not mutually exclusive; both mechanisms might have contributed to the difficulty that listeners experienced in the speaker incongruent condition. We will reconsider these two mechanisms in light of the new results presented below.

As mentioned above, there have been numerous investigations of sentence context effects on lexical ambiguity resolution, many using spoken stimuli, but to our knowledge, no prior investigations have manipulated the social characteristics of the speaker's voice. Nygaard and

Lunders (2002) comes the closest; they examined the effect of vocal affect on the transcription of heterographic homophones. The authors selected homophones like “ate/eight” that had one happy meaning and one neutral meaning, like “blue/blew” that had one sad meaning and one neutral meaning, and like “chews/choose” that had two neutral meanings. The emotional valence of the words was assessed by a separate rating study. In the primary experiments, the spoken homophones were presented in isolation, while varying emotional tone of voice (happy, neutral, or sad). Vocal affect influenced how the homophones were transcribed, with more happy transcriptions for words spoken in a happy voice, and more sad transcriptions of words spoken in a sad voice. The Nygaard and Lunders study demonstrated that voice quality provides a type of context that guides lexical disambiguation; However, the transcription task doesn’t allow us to draw inferences about how vocal quality constrained the initial activation of the homophonic words. Thus, it is not possible to compare the timecourse of the voice quality effect with the timecourse of sentence context effects, nor to make strong inferences about the cognitive mechanisms supporting the voice effect.

The Current Study

We conducted three experiments designed to investigate how social qualities indexed by the voice influence lexical ambiguity resolution. We used a visual world eye-tracking paradigm, similar to Chen and Boland (2008), to test specific predictions about the time-course of speaker congruity effects. For this project, we selected 24 homophone meanings that had a stereotyped bias (assessed by a questionnaire, described below) on either the male-female dimension or the adult-child dimension. To create the spoken stimuli, we recruited four speakers, each exemplifying one end of these two dimensions. The same 24 homophones, four voices, and visual stimuli were used in all three experiments. For Experiment 1, the homophones were

embedded in biasing sentence contexts, such as (1) above and participants engaged in a passive listening task. In Experiments 2 and 3, the homophones were embedded in a neutral "Look at the…" sentence context, creating a directed looking task. The difficulty of the directed looking task was manipulated by adding crowd noise to the speech signal for Experiment 3, making word recognition more difficult compared with Experiment 2, which had no added background noise.

In all three of our experiments, the participants saw only a central fixation cross at the start of each trial. At the onset of the spoken homophone, four images appeared: one representing a homophone meaning, one representing a phonological competitor of the homophone, and two unrelated images. The primary dependent measure was the latency of the first fixation on the homophone picture.

We expected that the voices of our four speakers would activate stereotyped information about the social categories they belong to, along the dimensions of male-female and adult-child. We further hypothesized that such social stereotypes would impact the activation of alternative homophone meanings. The important linking assumption is that the picture representing a homophone meaning should be fixated more rapidly when the corresponding homophone sense is accessed more rapidly and/or more strongly. For example, if the money sense of the word "check" is more accessible when uttered by an adult as opposed to a young child, an image of such a check should be fixated more rapidly in response to the adult utterance compared with the child utterance.

The exemplar model account of Hay and Walker (2011) predicts voice congruity effects across all three experiments. Regardless of sentence context or task difficulty, it should be easier to recognize a word when it is uttered by a voice acoustically similar to the predominant voices in one's prior experience of the word. This prediction rests on one additional assumption, that the

association norms (described below) that we collected to evaluate the social biases of our homophone provide good estimates of our prior social experiences of the words.

A theoretical model in which speaker congruity effects are relatively slow to emerge, such as that supported by McLennan and Luce (2005), does not predict congruity effects across the board. Such a model predicts speaker congruity effects only when fixation times are relatively slow and perceptual processing is relatively effortful. For our experiments, that means larger speaker congruity effects in Experiments 1 and 3, compared with Experiment 2.

Finally, if social congruity effects emerge as a result of sentence-level anticipatory processes, speaker congruity effects should be strongest in Experiment 1, and perhaps strongest of all in those sentences with self-referential pronouns. This pattern is predicted because the social characteristics of the speaker are most relevant in first-person narrative contexts. In linguistically neutral contexts like "Look at the…," the continuation of the sentence is so unconstrained as to make anticipatory processing fruitless. Anticipatory processes play an important role in current theories of sentence processing (e.g., Altmann & Kamide, 2007; Boland, 2005, Crocker & Brant, 2001; DeLong et al., 2005; Frisson et al., 2005; Gibson, 1998; Hale, 2003; Jurafsky, 1996; McDonald & Shillcock, 2003, van Berkum et al., 2005; see Kamide, 2008, for a recent review).

Within the larger set of theories that incorporate predictive processing, constraint-based lexicalist theories are especially relevant because they allow for non-linguistic, situation-specific cues to guide ambiguity resolution. Up to now, such theories have not explicitly considered social inferences based on vocal quality as a cue for ambiguity resolution, but other types of non-linguistic situational contexts have been found to play a role in syntactic ambiguity resolution (e.g., Chambers et al., 2004; Tanenhaus et al., 1995), lexical ambiguity resolution (e.g.,

Halberstadt et al., 1995; Nygaard & Lunders, 2002), and referential ambiguity resolution (e.g., Clark et al., 1983; Hanna et al., 2003; Sedivy et al., 1999). Thus, under a constraint-based account of sentence comprehension, it is plausible that voice-based social cues would cause listeners to develop expectations that would guide ambiguity resolution. In fact, in the context of a relative clause attachment ambiguity, Kamide (2012) demonstrated that listeners learn a speaker's attachment preference and anticipate different attachment behavior from different speakers during the course of an experiment, based solely on their recorded voices, i.e., the speakers were not present during the experiment.

Top-down context effects have played an important role with the lexical ambiguity resolution literature, with current theories allowing constraining context to guide lexical ambiguity resolution to some degree. Thus, the selective access model could be easily modified to include social inferences based on vocal quality as part of the context (e.g., Simpson & Krueger, 1991; Tabossi, 1988). Selective access theories allow for sentence context (and presumably the speaker's voice) to facilitate access to both dominant and subordinate meanings and easily accommodate graded effects. Thus, we assume that the biasing contexts in Experiment 1 would be made slightly strongly by a socially congruent voice and slightly weaker by an incongruent voice, predicting a main effect of congruency. This could also lead to congruency effects with the neutral contexts in Experiments 2 and 3, though as argued above, even if the voices biased the contexts slightly, the contexts would still not be very constraining.

The reordered access model posits that less frequent meanings of an ambiguous word are activated more quickly (and more effectively compete with frequent meanings) in supporting sentential context (e.g., Duffy et al., 1988; Sereno et al., 2004). If the speaker's voice were taken to provide part of the sentence context, it might strengthen a sentential context that was

otherwise not quite strong enough to reorder access of the subordinate meaning. However, the predictions would depend upon both the meaning frequency of the homophone and the strength of the biasing context, neither of which we were able to manipulate in the current set of experiments. Thus, the current data cannot distinguish between these two theories of lexical ambiguity resolution. Nonetheless, our data will provide some evidence to suggest how either theory might incorporate speaker effects.

Experiment 1

This experiment utilizes a visual world passive looking task (Altmann & Kamide, 1999) to investigate the impact of a socially-relevant voice manipulation on fixations to images representing one meaning of a homophone. We chose homophones that were strongly biased towards children, adults, men, or women in our norming studies (described below). To create our auditory stimuli, the homophone was embedded in a sentence context consistent with the biased meaning. The pictured homophone meaning was always congruent with the linguistic bias of the sentence. In the experiment, participants heard the sentences in either a voice that was consistent with the social bias of the homophone or an inconsistent voice. For each participant, half of the sentences were spoken in an inconsistent voice.

Given this design, the linguistic context was a reliable cue to homophone meaning, as were the visual images (which appeared at homophone onset). In contrast, the social properties of the voice were consistent with the intended meaning only half of the time. Thus, there was no experiment-internal motivation for participants to attend to the social features conveyed by the voice, for purposes of homophone disambiguation.

About half of our contexts were in the first person with self-referential pronouns. Motivated by van Berkum et al. (2008), we were interested in whether speaker congruity effects

would be limited to sentences with self-referential pronouns. Social properties of the voice are most relevant to predictions about upcoming material in sentences about the speaker herself. Consistent with this prediction, Creel (2010) found that both children and adults used acoustic cues to talker for first-person requests ("I want the square"), but not third-person requests ("Billy wants the square"). This was optimal in her experiment, because participants knew that the two talkers preferred objects of different colors. Participants looked to objects of the preferred color only for first-person utterances.

Method

Participants. Thirty undergraduates from University of Michigan Introductory Psychology subject pool provided the eye-tracking data for the primary experiment and received partial course credit for participating. One of these participants was dropped for purposes of analysis due to the experimenter's comment that the participant laughed excessively throughout the experiment, combined with an unusually high level of missing data. A total of 93 additional participants from the subject pool and 42 additional paid participants from Mechanical Turk provided data for the norming studies described in the Materials section.

Materials. The stimuli used for this experiment include our set of 24 homophones, the carrier sentences that biased the listener toward the pictured homophone meaning, and the four images used for each experimental trial. In this section, we also describe the normative data we collected to assess the social associations for our four voices, the social biases of our homophones, the bias of our carrier sentences toward the intended homophone meaning, and the labels generated to describe our visual images.

Voices. Four voices were selected for use in this experiment. The male and female voices were provided by college students. The adult voice was a 48 year old woman and the child voice

was a 10 year old boy. Speakers were recorded in a sound-attenuating chamber, using a DAT recorder. The spoken stimuli for all three experiments were recorded in a single session, for each speaker. Twenty-two University of Michigan students who did not participate in the primary experiment rated three samples of each voice on four 7-point scales. The same three sentences were used for all four speakers: “Look at the club”, “Look at the top”, and “Look at the bow.” For each scale, the participant was asked “How masculine (or feminine or childlike or adultlike) does this voice sound?” Mean ratings are provided in Table 1. Higher values indicate greater perceived masculinity, femininity, and so forth.

------insert Table 1 about here----

Social Bias of Homophone Senses. A survey was administered via Mechanical Turk to 20 participants. All participants used IP addresses in the U.S. to respond to the survey and were self-reported to be at least 18 years of age. Participants were asked to rate word senses on three dimensions: male-female, adult-child, and upper class-lower class. Each endpoint of the dimension was queried individually (see instructions below). Our rating scales were modeled on those used by Nygaard and Lunders (2002). The male-female list included two meanings for each of 43 homophones. Homographic homophone senses were identified by an additional word, in parenthesis, e.g., “nail (finger).” The list also included 12 filler word senses that we judged to be polarized on this dimension, e.g. “mother” and “football (sport)”. The 98 items were randomly ordered and divided into blocks of 14-18 items that were presented together on a single page of the survey, with the constraint that two senses of a homophone could not appear together within the same block. Participants rated all words on a single scale, before evaluating the words on a different scale. The instructions for the male scale were as follows. *On a scale of 1-7, please rate how strongly you would associate the following words with MEN. The word in*

(parenthesis) explains more about the initial word. You are rating the first word. To form the other five scales, the word MEN was replaced with WOMEN, CHILDREN, ADULTS, UPPER CLASS, and LOWER CLASS, respectively. The child-adult list was contained 112 homophone pairs, plus 20 filler words. The upper-lower class list contained 27 homophone pairs and 26 filler words. Some items appeared on more than one list.

Twenty-four homophone senses were selected for use in the experiment, from either the male-female list or the adult-child list. These homophones can be found in the Appendix, along with the carrier sentences used for this experiment. For a few homophones (e.g., nails, sale/sail), two meanings met our criteria for selection, but for most homophones only one meaning was used for the experiment. Unfortunately, it was not possible to balance items equally across our four social categories of interest. We used two criteria to select our set of 24 homophone senses. First, the homophone sense had to be sufficiently imageable that we could find a photograph that passed the criteria for our labeling norms (see below). Second, the homophone sense had to have a strong social bias. We defined this as having a mean rating of 4.0 or higher on the biased end of the dimension, and 3.25 or lower on the unbiased end of the dimension, with a difference of at least 1.5. For example, the money sense of “check” had a mean rating of 5.82 on the adult scale and 1.1 on the child scale. Across all 24 items, the average rating on the biased scale was 5.61 and the average rating on the unbiased scale was 2.05, with an average difference of 3.56.

Because we had to control for a range of other factors, we were not able to control for meaning frequency. However, we estimate that the homophone meanings were about equally divided between dominant and subordinate meanings.

Sentence Contexts. For each homophone, we designed a sentence context to support the pictured image. The text of our sentence stimuli can be found in the Appendix. In most cases, the

sentence context did not completely rule out alternative meanings. To evaluate the degree to which our sentences constrained homophone meaning, we created a questionnaire using written versions of our 24 experimental context sentences and 10 similar sentences. Each item on the questionnaire consisted of the sentence context, ending with "…" and 7 point likert scales next to the two homophone meanings that we judged to be most common. An example is provided below in (5). Participants were asked to consider the naturalness of each completion independently and were given an example of a pair in which both meanings were rated as good completions. At the beginning of the questionnaire, and again on the top of each page, participants were told that 1 meant "awful" and 7 meant "great." Thirty-eight University of Michigan students completed the questionaire. The critical contexts were rated 6.67 on average, demonstrating that completion with the intended homophone meaning was very natural.

5. *When I feel creative, I improvise on my ...*

sax	1	2	3	4	5	6	7
sacks	1	2	3	4	5	6	7

Images. Twenty-four photographic images were selected to represent the intended sense of the homophone for each of our critical items. Similar images were selected to represent a phonological competitor and two phonologically and semantically unrelated words for each trial. Phonological competitors were a mix words that rhymed with the target word and words that shared at least two onset phonemes with the target word. Many pictures were used on more than one trial, with no picture occurring more than twice during the experiment.

Labeling norms were collected in two waves using Qualtrics surveys. Twenty-two Mturk participants participated in the first wave. Participants were shown a superset of potential images and asked to provide a one-word label. Candidate images for homophone and phonological competitors were discarded if there was not a clear majority of responses with the intended label.

Candidates for filler pictures were carefully screened for phonologically related or semantically related labels, and were not matched with homophone pictures unless the filler picture was semantically and phonologically unrelated. The second wave of labeling norms included the pictures that passed the first set, plus additional candidate pictures. Thirty-three University of Michigan participants labeled 113 images for these norms. Only pictures receiving more than 50% correct identifications were used in the experiment. The final set of images were given the intended labels 82% of the time for child-biased images, 92% of the time for adult-biased images, 87% of the time for male-biased images, and 86% of the time for woman-biased images.

Procedure. Before the start of the experiment, we collected written informed consent. We collected the eye movement data using the SMI Eylink II head-mounted eye-tracker using a 250 hz sampling rate. Participants were seated in front of the computer screen that presented the images and the head-tracker was placed on their head and adjusted to provide a comfortable, but firm fit. Two eye cameras were adjusted to provide an optimal eye image and then standard calibration and validation routines were run until a satisfactory calibration was achieved. In most instances, both eyes were calibrated and validated, but the best eye was used. A drift correction preceded each trial.

There was no explicit task other than listening to the sentences and looking at the pictures. Participants read written instructions after the calibration and validation procedures. They were told that that they would hear a sentence and four pictures would appear as they heard the final word of the sentence. One of the pictures would match the word they were hearing. They were told to look at the matching picture first, but could examine the other pictures if they wished after they looked at the matching picture. There were two practice trials before the experiment began.

For each trial, a central fixation cross appeared at the outset and remained on the screen while the audio file started playing. At the onset of the spoken homophone, a picture appeared in each of the four screen quadrants. Trials ended when the target image was fixated. Thus fixations to the other pictures were recorded only if they preceded looks to the target picture.

Across trials, the homophone referent image and phonological competitor image were equally likely to appear in each of the four possible screen locations. For each item, picture locations were identical in the congruent and incongruent conditions. Two experimental lists were created, such that an item that was in the congruent condition on List A was in the incongruent condition on List B. Half of the trials on each list were congruent and half were incongruent. The visual stimuli and the lexical content of the auditory sentence for congruent and incongruent conditions were identical; the only difference was the congruity of the voice in the audio file. The 24 experimental trials occurred in a different random order for each participant.

Results

Our primary dependent variable is the latency of the first fixation on the target pictures. For our analyses, we ignored fixations that occurred either less than 200 ms after homophone onset or more than 2000 ms after homophone onset. The target image was never fixated in less than 200 ms, but there were 13 trials (2%) with first target fixations of greater than 2000 ms which were removed prior to the analysis. There were also some trials on which the target was never fixated, especially in the incongruent condition. As a result, 96% of congruent trials and 85% of incongruent trials had a fixation on the target image within 2000 ms of homophone onset. This difference in fixation rate was significant by both participants and items in two-tailed paired t-tests ($\alpha = .05$). The average duration of the first fixation on the target was 94 ms (94.5 ms in the congruent condition and 93.5 ms in the incongruent condition).

The mean **latency of first target fixation** was 1091ms in the congruent condition and 1141ms in the incongruent condition. This speaker congruity effect was significant by participants and by items in two-tailed, paired t-tests (α =.05). The latency data were further analyzed using linear mixed effects models in R, with the lme4 package. We constructed maximal random-effects models, following Barr et al. (2013). Factors were sum-coded, producing the following model: latency ~ congruency.f * bias.f + (1 + bias.f | homophone) + (1 + bias.f | participants). Congruency and homophone bias were fixed factors; homophone and participant were random factors. We observed a main effect of congruency, with shorter target fixation latencies on trials with congruent voices, and an interaction of congruency and bias, with the largest congruency advantage for the male-biased homophones (101ms, N = 8) and child-biased homophones (126ms, N = 8), but no congruency effects for the female-biased homophones (-19ms, N = 6) or the adult-biased homophones (-122ms, N = 2). These effects are summarized in the upper panel of Table 2. The anova function was used on the fitted model to generate F statistics to evaluate the significance of the effects, at α = .05.

----insert Table 2 about here---

Due to trials without fixations, the speaker congruity effect appeared slightly larger when computed from the item means, with the congruent item means 61ms faster than the incongruent item means, on average. For the 15 items using self-referential, first-person pronouns, this difference dropped to 51ms, while the remaining items had an average 77ms advantage in the

congruent condition.[1] Thus, we found no evidence that the speaker congruity effect was dependent upon self-referential pronouns.

We also evaluated the hypothesis **that the fixation pattern across time** might differ for the congruent and incongruent conditions. We divided the critical region into bins of 100ms for each participant by congruency cell. For this analysis, we ignored homophone bias type, so that each cell would contain 12 trials. In each bin, we counted the proportion of trials for which there was a fixation, separately for each of the three image types (target, phonological competitor, two fillers). The fixation patterns for all three image types are illustrated in Figure 1.

---insert Figure 1 about here---

First, we focused on fixations to the target image in the congruent and incongruent conditions. We evaluated the difference between the two curves from 500ms to 1000ms after word onset with a growth curve analysis, following Mirman et al. (2008). We selected this range of bins in order to assess fixation patterns from the time participants first began fixating the images until the initial target fixations peaked (see Fig 1). We modelled the time course of target fixations with a third-order (cubic) orthogonal polynomial and fixed effects of congruency (within-participants) on all time terms. The model also included participant random effects on all time terms and participant-by-congruency random effects on all time terms except the cubic (because it tends to capture less-relevant effects in the tails). The results are reported in Table 3, with the polynomial terms labeled *ot1* to *ot3*. The fit of the model to the observed data is illustrated in Figure 2. There was a significant effect of congruency on the intercept term

[1] Items with self-referential pronouns constituted at least half of each bias type: 50% of female items, 63% of child-biased and male-biased items, and 100% of adult-biased items.

(labelled *Condition nc*), indicating lower overall target fixation proportions for the non-congruent condition relative to the congruent condition. However, congruency did not interact with the polynomial terms, suggesting the same general shape in the growth curve of target fixations in the congruent and incongruent conditions.

----insert Table 3 & Figure 2 about here----

As shown in Figure 1, target fixations did not appear to outnumber looks to the phonological competitor and filler pictures until about 900ms after homophone onset in the incongruent condition, though the target image may have been preferred slightly earlier in the congruent condition. We evaluated this potential difference by converting the proportions for all three image types to log odds. Two-tailed paired comparisons ($\alpha = .05$) in the 800ms bin found that the congruent targets were fixated more often than the fillers, though not more often than phonological competitors, whereas the incongruent targets did not differ from either the fillers or the phonological competitors. By the 900ms bin, both congruent and incongruent targets were fixated more often than their respective fillers and phonological competitors.

Discussion

Participants were faster and more likely to fixate the homophone image when the sentence was spoken in a socially consistent voice. Fixating the homophone image indicates that participants have identified the referent of the homophone, which itself requires lexical ambiguity disambiguation. We observed a small, but robust, effect of speaker congruity, looking at the same phenomena from four different perspectives: the probability of a fixtion within 2000ms, the latency of the first fixation, the likelihood of fixating the target over time, and the time bin at which target fixations start to outnumber other fixations.

The consistent effect of speaker congruity is especially striking, given that the homophone meaning was disambiguated by the linguistic context. Note that, within the context of our experiment, the linguistic bias of the sentence was much more reliable than the social cues associated with the voice. That is, the homophone image was always consistent with the linguistic bias of the sentence, but the homophone was only consistent with the voice on half of the experimental trials. While we can't tell from this experiment whether the relationship between vocal cues and linguistic cues are additive, both types of cues appear to have influenced processing simultaneously.

In fact, this finding is not unexpected given van Berkum et al.'s (2008) ERP study, in which they found a small N400 for unambiguous words in a narrative sentence context, when the word was incongruent with the speaker's voice, e.g., "Every evening I drink some **wine** before I go to sleep," spoken by a young child. The current finding demonstrates that speaker congruity effects can be found with visual world eye-tracking, and furthermore, that speaker congruity effects can be found for homophones in disambiguating context.

The interaction of congruency with bias type in the latency analysis indicates that our stimuli were not equally effective at producing the speaker congruity effect. We attribute this to the limitations of our necessarily small set of homophones. Unfortunately, we were limited by the number of homophones that passed our norms for both social biases and imageability. In the two bias categories where there didn't seem to be any (positive) speaker congruity effect, there were only eight items combined, and three of these had high numbers of missing values in the non-congruent condition because listeners did not look at the target image during the critical interval. Across all eight items in the female and adult bias categories, the target image was fixated 95% of the time in the congruent condition, but only 67% of the time in the incongruent

condition (vs. 96% congruent fixations and 85% incongruent fixations in the full dataset). Thus, there is evidence of a speaker congruity effect for female and adult bias items in the probability of fixation, even though it was absent in the latency data.

Our findings can be explained by a predictive account, which can also explain the findings of van Berkum et al. (2008). Under such an account, listeners used both the lexical content of the sentence and the social properties associated with the speaker's voice to anticipate what the speaker was likely to say. Slower fixation times in our study, and an N400 effect in van Berkum et al.'s study, indicate that the continuation of the sentence was inconsistent with the listeners' expectations. Note, however, that listeners' eye movements did not reflect the difference in predictive capacity between 1st-person narratives and non-1st-person narratives.

An alternative account, which explains these findings equally well, is motivated by exemplar theories of word recognition, such as that advocated by Hay and Walker (2011). On their account, the source of the speaker congruity effect is bottom-up lexical access, not top-down predictions. Under the exemplar theory, auditory lexical access is most efficient when a word is uttered in a voice that is acoustically (and perhaps socially) similar to our previous experiences of that word. For example, "nails" in a female voice might more readily activate the fingernails meaning if we have heard women using "nails" to mean fingernails more often than we've heard men using "nails" to mean fingernails. Crucially, such bottom-up effects ought to be independent of the carrier sentence. Thus, if the speaker congruity effect in Experiment 1 is due to exemplar similarity, then we ought to find a similar speaker congruity effect in Experiments 2 and 3, using a completely unconstrained sentence context.

In Experiments 2 and 3, every sentence was of the form "Look at the X," with the images appearing at the onset of the noun that replaced X. In such an experimental context, listeners

cannot predict what X is likely to be. To be sure, listeners might still use vocal cues to anticipate categories of things that men would be likely to say, or that women would be likely to say. However, such categories would be quite broad and distinct from the more specific expectations that could be developed during the constraining sentence contexts of Experiment 1.

Experiment 2

In Experiment 2, we embedded the homophones in a completely neutral linguistic context, e.g., “Look at the nails.” Carrier phrases such as “look at” or “click on,” which direct the listener to do something, are often used in speech perception experiments where the goal is to examine the time-course of spoken word recognition (e.g.,Beddor et al., 2013; Dahane et al., 2001; Gow & McMurray, 2007). This type of directed action task facilitates a tight link between perceptual processing of the spoken word and eye fixations on a visual representation of the word, allowing researchers to investigate the use of detailed phonetic information. For example, Beddor et al. manipulated the onset of coarticulatory nasalization in the vowel of a word like “bend” and found that listeners fixated the target picture more rapidly when coarticulation began earlier. For our purposes, this paradigm should be maximally sensitive to the use of bottom-up perceptual cues, while eliminating most top-down cues that might influence word recognition.

An important difference between our experiment and the speech perception experiments is that we did not allow participants to preview the pictures. As in Experiment 1, the pictures appeared at the onset of the spoken homophone. Under such circumstances, will the speaker congruity effect be bigger, smaller, or the same, compared with Experiment 1? The exemplar account predicts that the speaker congruity effect should be similar across the two experiments, because the bottom up advantage afforded by the congruent voice is independent of the carrier

sentence. If the speaker congruity effects in Experiment 1 were caused by the greater acoustic similarity of speaker-congruent tokens to stored exemplars, a robust speaker congruity effect should be found in Experiment 2 as well. In contrast, the predictive account relies upon predictions generated during the carrier phrase. Much more specific predictions can be generated from the narrative sentences used in Experiment 1 compared with the neutral carrier phrases in Experiment 2. On one hand, this could make predictive processing less likely to occur at all in Experiment 2, because trying to anticipate the last word in the sentence is futile. If so, little or no effect of voice congruity should be observed in Experiment 2. On the other hand, the four different voices are the only thing that differs among all the carrier phrases heard in Experiment 2. This could potentially amplify attention on the voices and their social properties, thereby increasing the size of the voice congruity effect.

Method

Participants. Thirty University of Michigan students from the Introductory Psychology subject pool received partial course credit for participating in the experiment. We omitted the last participant run on one of the lists, in order to have exactly 15 participants on each list.

Materials. The voices, homophones, and images were identical to those used in Experiment 1. However, for this experiment, the same carrier phrase, “Look at the…,” was used on every trial. Our four speakers were recorded producing complete sentences, i.e., “Look at the bat.” However, for experimental trials, the utterances were spliced at the onset of the homophone. A single token of “Look at the” was selected for each speaker, based on the clarity of the speech, especially regarding the absence of coarticulation between “the” and the following homophone. This audio file was used as the sentence context for all of that speaker’s trials. The homophone was presented in a separate audio file. As in Experiment 1, presentation of the images coincided

with the onset of the homophone and the only difference between the congruent and incongruent trials was the social congruity of the voice with the pictured meaning of the homophone.

Procedure. In contrast with Experiment 1, the sentence context itself instructed the listener to look at the target picture. Thus the participant's understanding of the task was different than in Experiment 1, which used a passive looking paradigm. Otherwise, the procedures and equipment were identical to those used in Experiment 1, except that each participant received a random 22-item subset of the 24 experimental trials, due to a programming error. After the calibration and validation procedures, the participants read written instructions. They were told that four pictures would be displayed as they heard instructions about which picture to look at. Participants were instructed not to move their eyes until they heard the picture label, but to look as the appropriate picture as soon as they recognized the word.

Results

In this experiment, every trial included a fixation on the homophone image. As in Experiment 1, we removed fixations that occurred less than 200ms after homophone onset as well as those that occurred more than 2000ms after homophone onset. This excluded 2% of the trials in each of the congruent and the incongruent conditions. Thus, for our purposes, the probability of a first fixation on the target picture during the critical window was 98% in each condition. The mean duration of the first fixation on the homophone image was 491 in the congruent condition and 475 in the incongruent condition, considerably longer than in Experiment 1.

The average **latency of the first fixation** on the homophone image was 759ms in the congruent condition and 739ms in the incongruent condition. This difference was not statistically significant, neither by paired t-tests on item and participant means ($\alpha = .05$), nor in a linear

mixed model, using the same model as in Experiment 1 and close variations [Fs < 1, see middle panel of Table 2].

As in Experiment 1, we also evaluated the hypothesis **that the fixation patterns across time** might differ for the congruent and incongruent conditions. The fixation pattern for the target pictures is illustrated in Figure 3 and the looking pattern for all three types of pictures is illustrated in Figure 4. The change in the shape of the curves, compared with Figures 1 and 2, reflect the substantially longer target fixations observed in Experiment 2, due to the directed looking task. We analyzed the target fixation patterns using the same growth curve analysis methods as in Experiment 1. We included the temporal bins from 200ms to 1200ms in our analysis, based on a visual inspection of the data-points in Figure 3. Impressionistically, the congruent and incongruent proportions are very similar over the first 800ms after homophone onset, with any advantage falling to the incongruent condition. It is only when target fixations start to decrease (around 900ms) that fixations in the congruent condition seem to outnumber those in the incongruent condition, as if participants remained fixated on the target longer in the congruent condition. Even so, we found no evidence of an advantage for the congruent voice, consistent with the fixation latency results from the current experiment, but in contrast with Experiment 1 Nor did any of the polynomial terms interact with congruency, suggesting that the fixation patterns across time showed the same pattern for congruent and incongruent conditions. See Table 4.

---insert Table 4 and Figures 3 & 4 about here---

Figure 4 illustrates how quickly after homophone onset target fixations became more common than looks to the phonological competitor and filler picture, providing evidence that the

homophone has been recognized and disambiguated. The figures suggest that participants had identified the target picture as the referent of the pronoun by around 500ms after homophone onset in both conditions. There is no evidence that referent identification occurred earlier in the congruent voice condition than the incongruent voice condition; if anything, the opposite is true.

Discussion

Whereas a speaker congruity effect was consistently found in Experiment 1 across four different measures, there was no evidence of a speaker congruity effect in Experiment 2 on any of the four measures. This is remarkable, because the directed action task used in Experiment 2 more closely time-locks eye fixations to the speech signal, usually resulting in greater task sensitivity. The total absence of a speaker congruity effect argues against a bottom-up exemplar account of the speaker congruity effect in Experiment 1. Instead, it suggests that the narrative carrier sentences in Experiment 1 encouraged predictive processing, and that social characteristics of the voice modulated those predictions.

Compared with Experiment 1, fixations on the homophone image occurred much more rapidly and were of much longer duration. This difference reflects our shift to a directed looking task from the passive listening paradigm used in Experiment 1. The difference between Experiments is particularly striking in the first 600 ms: in Experiment 2, participants began fixating the images within 200-300 ms of homophone onset, but in Experiment 1, they did not begin looking at the pictures until 600-700 ms after homophone onset.

Because we altered both the nature of the carrier sentence and the nature of the task between Experiments 1 and 2, we cannot conclude that the speaker congruity effects in Experiment 1 were due to predictive processing triggered by the narrative carrier sentences without additional evidence. An alternative possibility is that the speaker congruity effects

emerge only when processing is relatively difficult. As noted above, McLennan and Luce (2005) found talker-specificity effects for a difficult lexical decision task, but not for an easy lexical decision task. They reasoned that talker-specific effects were relatively slow to arise and only emerged when the task was challenging and response times were slow. Similar reasoning is offered by Sumner et al. (2014). The fact that target fixations occurred more rapidly and were more likely in Experiment 2 suggests that the task was more effortless than in Experiment 1. In Experiment 3, we will explore this possibility by adding noise to the "Look at the X" stimuli from Experiment 2. The goal is to make the directed looking task more difficult and to slow down the target fixations in order to allow more time for the speaker congruity effects to emerge.

We also considered the possibility that our failure to find a voice congruity effect might be due to the gender distribution in our participant pool. Van den Brink et al. (2012) found that the speaker congruity N400 effect was modulated by gender, while offline judgments about the oddity of speaker-incongruous sentences were not modulated by gender. In their ERP experiment, female listeners showed a larger speaker congruity N400, compared with male listeners. This gender effect in the ERP response was explained by differences in empathy, as measured by Baron-Cohen & Wheelwright's (2004) Empathy Quotient questionnaire. However, a post hoc analysis of our data revealed no effect of participant gender. Neither men nor women showed any trend toward a speaker congruity effect in Experiment 2.

Experiment 3

The goal of Experiment 3 was to make processing more difficult in the directed looking task, in order to allow time for speaker congruity effects to emerge. This logic is based on the McLennan and Luce (2005) finding that talker-specific effects in an auditory lexical decision task was observed when the non-words were word-like, but not when the non-words were easy to

discriminate phonologically. The difference in task difficulty resulted in a 30-40 ms difference in response time, which was sufficient to allow the talker-specificity effects to emerge in the more difficult version of the task. They also reported a shadowing paradigm in which participants repeated the spoken word as soon as possible or following a response cue given 150 ms after word offset. Mean response latencies were 808-855 from word onset in the immediate condition (no talker-specificity effects) and 350 – 388 ms from cue onset in the delayed condition (talker specificity effects observed). Word durations averaged about 564 ms, making the actual latency difference between the immediate and delayed conditions about 247 ms in the control condition. Together, these findings suggest that the emergence of speaker-dependent effects depends upon the time-course of processing across a variety of spoken language paradigms

The difference in mean fixation latencies on the target image between our Experiments 1 and 2 was 367 ms--even greater than the difference in shadowing latencies for McLennan and Luce (2005) and much greater than their difference in lexical decision latencies. However, it is difficult to draw conclusions based on the different time-course of fixations in our Experiments 1 and 2 because Experiment 1 employed a passive listening paradigm while Experiment 2 employed a directed looking task. Experiment 3 is designed to be more directly comparable with Experiment 2. The only difference between Experiments 2 and 3 is the addition of noise to the speech files, in order to make perceptual processing of the homophone more difficult, analogous to McLennan and Luce's nonword manipulation in their auditory lexical decision experiments.

Method

Participants. Thirty undergraduates from University of Michigan Introductory Psychology subject pool provided the eye-tracking data for the primary experiment and received

partial course credit for participating. Two additional participants completed the experiment, but their data were not included due to inaccurate tracking of their eye movements.

Materials. The images and homophones were identical to those used in the prior experiments. The speech files from Experiment 2 were overlaid with crowd noise, using the following procedure. First several minutes of crowd noise ("Crowd Talking 1," recorded prior to the start of a concert) were downloaded from soundjay.com. Second, all auditory stimulus files, including the crowd noise, were scaled in intensity to 70 db SPL. Recall from Experiment 2 that the same "look at the" context was used on every trial for a given speaker, with the homophone in a separate auditory file. Thus, an arbitrary splice point was chosen in the crowd noise file, such that we had a context portion that ended at the splice and a homophone portion that began at the splice. Finally, for each of the context speech files and the homophone speech files, a segment of crowd noise equal to the duration of the speech file plus 50ms (appended at the end of the speech file) were mixed with the speech files into a single mono track.

Procedure. The procedure was identical to that for Experiment 2, except for one sentence added to the written instructions: "The audio instructions are meant to mimic a natural setting like a busy food court or a crowd prior to a concert."

Results

Mean fixation latencies were computed, after removing outliers as in Experiments 1 and 2. Outliers accounted for 6.5% of the data, but were more common in the congruent condition than the incongruent condition ($p < .01$, 2-tailed t-tests, both by participants and by items). This difference suggests that participants actually had more difficulty identifying the referent of the homophone in the congruent condition. Mean fixation latency to the homophone image was

787ms in the congruent condition and 814ms in the incongruent condition. This difference was not significant in two-tailed t-tests on the subject and item means. A linear mixed model matching that used for prior experiments found no main effect of congruency, nor an interaction between congruency and bias (see lower panel of Table 2, above).

The fixation patterns over time were binned as in prior experiments. Figure 5 summarizes the target looks and Figure 6 summarizes the looks to all three image types. Incongruent and Congruent target fixations from 200ms to 1200ms were analyzed using the same growth curve analysis technique used in prior experiments. If we compare Figure 5 to Figure 3 (from Experiment 2), the overall shape of the curve is very similar, but without the congruency advantage as target fixations begin to decrease. The peak of the curve also appears to be lower in Experiment 3 than it was in Experiment 2. As in Experiment 2, there was no main effect of congruency, nor any interactions (see Table 5).

---insert Table 5 & Figure 5 about here---

As in Experiment 2, fixations on the incongruent target appear to outnumber those on the fillers and phonological competitors earlier than for the congruent condition, on approximately the same time-course (compare Fig 6a to Fig 6b). Thus, there was no evidence that hearing the homophone in a congruent voice led participants to preferentially fixate the homophone image earlier than when the homophone was heard in an incongruent voice. In contrast to Experiments 1 and 2, there does appear to be an advantage for phonological competitors compared with fillers, probably due to the addition of noise.

---insert Figure 6 about here---

Finally, we analyzed the latency data from Experiments 2 and 3 together to examine the effect of noise, if any, on congruency and homophone bias. A linear mixed model analysis combined the latency data from Experiments 2 and 3, using participant and item as random effects, and noise level (i.e. experiment), congruency, and homophone bias as fixed effects. The crucial congruency by noise interaction was not significant, nor was the three-way interaction (see Table 6). As expected, there was a main effect of noise, with faster fixations on the target images in Experiment 2 compared with Experiment 3. These results indicate that fixating the target image was more difficult in Experiment 3 than in Experiment 2, but that the difficulty level did not impact the effect of either speaker congruency or homophone bias. Importantly, even with the greater statistical power acquired by combining the data from the two experiments, no main effect of congruency and nor an interaction of congruency with homophone bias was observed.

---insert Table 6 about here---

Discussion

There was no evidence of a speaker congruity effect in any of the four measures we examined in Experiment 3, nor in the combined data from Experiments 2 and 3. Fixation latencies averaged about 50ms slower in Experiment 3, compared with Experiment 2, suggesting that our noise manipulation did indeed make the task more difficult. This latency difference was comparable to that obtained by McLennan and Luce (2005) in their auditory lexical decision experiments. However, in contrast to McLennan and Luce, we found that the increased difficulty in our study did not cause an effect of speaker voice to emerge.

The absence of speaker congruity effect would be surprising if the congruity effect observed in Experiment 1 were due to a bottom-up effect of the type advocated in Hay and

Walker (2008). However, it is consistent with the predictive processing account of the Experiment 1 effects.

General Discussion

Across three experiments, each designed to investigate how lexical ambiguity resolution is influenced by social characteristics indexed in the speaker's voice, we found evidence of such an influence only in Experiment 1. Interestingly, Experiment 1 used narrative, biasing sentence contexts and a passive looking paradigm—giving it the most ecological validity and arguably the least sensitivity to subtle effects. It is also the experiment that most closely mirrors the ERP and fMRI experiments that found speaker congruity effects for unambiguous words (van Berkum et al., 2008; Tesink et al, 2009; van den Brink et al., 2012). Thus, it seems that listeners can and do use social cues indexed by the speaker's voice to guide expectations about upcoming linguistic content. The observed speaker congruity effects cannot be described as an adaptive strategy to the experiment itself, because during the experiments, the speaker's voice was congruent on half of the trials and incongruent on half of the trials. These findings are consistent with psycholinguistic theories that allow non-linguistic knowledge to guide anticipatory processing, such as constraint-based lexicalist models of sentence processing and selective access models of lexical ambiguity resolution. Further, these results highlight the speaker's voice as a potential source of predictive constraints.

On the other hand, the absence of speaker congruity effects in Experiments 2 and 3 suggests that listeners do not always use social cues indexed by the voice. Use of such cues may depend on both the linguistic context and the listener's task-related goals. The same speakers, homophones, and images were used in all three experiments, but no speakers congruity effects were observed in neutral "Look at the…" contexts, even when background noise was added to

make word recognition more difficult. We suspect that Experiments 2 and 3 discouraged predictive processing because the contexts were almost completely non-constraining. In such contexts, the bottom-up phonetic cues to word identity appeared to be solely or primarily driving fixations.

We don't doubt that certain changes to the experimental paradigm using "Look at the…" contexts would result in speaker congruity effects. For example, if we had allowed preview of the images, listeners might have fixated speaker-congruent images even prior to hearing the homophone. We didn't run that version of the study because we were most interested in how anticipatory processes affected homophone recognition as the word was enfolding in time. Another version of the experiment would present images of both a speaker-congruent meaning and a speaker-incongruent meaning of the homophone. We suspect that listeners would fixate the speaker-congruent images more often and earlier than speaker-incongruent images. However, it's not clear that such a finding would reflect anticipatory use of the speaker's voice to guide homophone resolution. Rather, listeners might access both pictured homophone meanings from the phonetic cues and then use the speaker's voice to resolve the dilemma of which image to fixate.

These findings do not rule out the bottom-up exemplar account of speaker congruity effects advocated by Hay and Walker (2011), but our findings make it very unlikely that such a bottom-up mechanism produced the speaker congruity in Experiment 1. If a bottom-up mechanism were responsible, similar speaker congruity effects should have been observed in Experiments 2 and 3. Nonetheless, bottom-up and top-down mechanisms for speaker congruity are compatible, in principle. The different mechanisms might predominate under different circumstances.

One factor that limits the ability of the current experiments to evaluate the exemplar account of speaker congruity is that homophone biases might not align with listener experience. According to Hay and Walker's (2011) account, we recognize words by comparing the current acoustic signal to stored memory traces for previous utterances. Thus, words (or homophone senses) most frequently uttered by children should be more easily recognized in a child's voice because the current token will be acoustically similar to the predominant stored tokens. Consider, however, that child speech represents a small proportion of one's overall linguistic experience, for most listeners. Based on raw frequency alone, we would expect that our participants heard our child-biased homophone meanings most often in adult voices, creating a mismatch between the subjective bias of the homophone meanings and the listener's actual experience. The dual-route account of Sumner et al. (2014) avoids this mismatch problem by claiming that it is the conceptual biases, not the raw frequencies, which are most relevant. As a result, Sumner et al.'s claim are more easily testable, since there is no sufficiently large corpora of spoken English that is coded in terms of the speaker's age and gender.

Despite the lack of an ideal corpus, we did attempt frequency counts for our male and female biased homophones, using the Soap Opera Corpus (100 million words), the Switchboard Corpus (3 million words), and the Michigan Corpus of Academic Speech (1.8 million words. For some of our words we found few or no items, but our best estimate is that our female-biased words were well-aligned with the offline norms, with 76% of tokens uttered by women. The male-biased words were much more balanced, with only 49% of tokens uttered by men. Note that this analysis rests upon the assumption that these corpora are representative, not only of our participants' exposure to the critical homophones, but also of the proportion of male and female adult speech encountered by our undergraduate participants. We did not do the analogous corpus

analysis for the child-biased and adult-biased items because we assumed that, even for our child-biased items (such as the toy meaning of “jacks”), our participants had encountered those senses most often in adult speech, simply because they have been exposed to much more adult speech than child speech.

Conclusion

We found an eye movement analog to the N400-like speaker incongruity effect previously observed in ERP and fMRI experiments (van Berkum et al., 2008; Tesink et al., 2009, van den Brink et al., 2012). The prior experiments always used narrative Dutch contexts and the Dutch target word was sentence medial. Our experiments extend that line of research by replicating the effect with a different paradigm, a different language, and sentence-final homophones as the socially-biased target words. More importantly, we have uncovered some important clues as to the cognitive mechanisms giving rise to the effect. While we replicated the speaker incongruity effect with narrative contexts and a passive looking task, we were not able to replicate the effect using a neutral “Look at the X” context and task. Our speaker congruity effect only occurred following constraining linguistic contexts, in which the listener could reasonably generate expectations about upcoming material. We think these top-down expectations are the primary mechanism for speaker congruity effects that arise during spoken word recognition. Our study differed from the previous studies in that the target words were always homophones with at least one socially-biased meaning. In this respect, our study was similar to Nygaard and Lunders (2002), who examined the recognition of emotionally-biased homophones with congruent and incongruent emotional voices. They also found congruity effects, although their task (transcription of spoken words) was less closely linked to the temporal unfolding of the spoken word.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, **73**, 247-264.

Altmann, G. T. M. & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory & Language, 57*, 502-518.

Baron-Cohen, S. & Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with Asperger syndrome or high functioning autism and normal sex differences. *Journal of Autism and Developmental Disorders, 34*, 163-175.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America, 133*, 2350-2366.

Boland, J. E. (2005). Visual arguments. *Cognition, 95*, 237–74.

Chambers, C. G.; Tanenhaus, M. K.; Magnuson, J. S. (2004). Actions and Affordances in Syntactic Ambiguity Resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 687-696. doi: 10.1037/0278-7393.30.3.687

Chen, L. & Boland, J. E. (2008). Dominance and context effects on activation of alternative homophone meanings. *Memory & Cognition, 36*, 1306-1323.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. J*ournal of Verbal Learning & Verbal Behavior, 22*, 245-258.

Creel, S. C. (2010). Considering the source: Preschoolers (and adults) use talker acoustics predictively and flexibly in on-line sentence processing. *Proceedings of the 32nd annual Cognitive Science Society Conference,* Portland, OR.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heading the voice of experience: The role of talker variation in lexical access. *Cognition, 108,* 633-664.

Creel, S. C. & Tumlin, J. A., (2011). On-line acoustic and semantic interpretation of talker information. J*ournal of Memory & Language, 65,* 264-285.

Crocker, M. W. & Brant, T. (2001). Wide coverage probabilistic sentence processing. *Journal of Psycholinguistic Research, 29,* 647-669.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., and Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition, *Language & Cognitive Processes, 16,* 507–534.

DeLong, K.A., Urbach, T.P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8,* 1117–1121.

Duffy, S., Morris, R., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language, 27,* 429-446.

Fine, A. B. and Jaeger, T. F. (2013). Syntactic priming in language comprehension allows linguistic expectations to converge on the statistics of the input. In Knauff, M., Pauen, N., Sebanz, & I. Wachsmuth (eds), *Proceedings of the 35th annual meeting of the Cognitive Science Society (CogSci13),* 3835-3840. Austin TX, Cognitive Science Society.

Frisson, S., Rayner, K., and Pickering, M. (2007). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 862-877.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68, 1–76.

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

Goldinger, S. D., Pisoni, D. B., & Luce, P. (1996). Speech perception and spoken word recognition: Research and theory. In N. J. Lass (ed)., *Principles of experimental phonetics,* 277-327, St. Louis, MO: Mosby Year-Book.

Gow, D. W. & McMurray, B. (2007). Word recognition and phonology: The case of English coronal place assimilation, in J. Cole and J. I. Hualde (eds), *Laboratory Phonology 9,* Mouton de Gruyter, Berlin, 173–200.

Halberstadt, J. B., Niedenthal P. M., & Kushner, J. (1995). Resolution of lexical ambiguity by emotional state. *Psychological Science, 6,* 278-282.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research,* 32, 101–123.

Hanna, J. E. Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory & Language, 49,* 43-61.

Johnson, K. (1997) Speech perception w/o speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (eds.), *Talker variability in Speech Processing,* 145-165, Morgan Kaufmann Publishers Inc, San Francisco, CA.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34,* 485-499.

Jurafsky, Daniel. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science,* 20. 137–94.

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language & Linguistic Compass, 2,* 647-670. DOI: 10.1111/j.1749-818X.2008.00072.x.

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition, 124,* 66–71.http://dx.doi.org/10.1016/j.cognition.2012.03.001.

Kamide, Y., Altmann, G.T.M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133–156. http://dx.doi.org/10.1016/S0749-596X(03)00023-8.

Klatt, D. H. (1989). Review of selected models of speech perception. In W. Marslen-Wilson (ed), Lexical representation and process, 169-226. Cambridge, MA, The MIT Press.

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

Ladefoged, P. & Broadbent, D. E. (1957). Information conveyed by vowels. Journal of the Acoustic Society of America, 29, 98-104.

Leinenger, M. & Rayner, K. (2013). Eye movements while reading biased homographs: Effects of prior encounter and biasing context on reducing the subordinate bias effect. *Journal of Cognitive Psychology, 25,* 665-681.

Martin, C., Vu, H., Kellas, G., & Metcalf, K. (1999). Strength of discourse context as a determinant of the subordinate bias effect. *Quarterly Journal of Experimental Psychology, 52A,* 813–839.

McDonald, S.A. & Shillcock, R.C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research, 43* 1735–1751.

McLennan, CT & Luce, PA (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31,* 306-321.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. Journal of Memory & Language, 59, 475-494.

Mullenix, J. & Pisoni, D. B. (1990). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365-378.

Nygaard, L.C. & Lunders, E.R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition, 30,* 583-593.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309-328.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N., (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition, 71,* 109-147.

Sereno, S. C., Brewer, C. C., & O'Donnell, P. J. (2003). Context effects in word recognition: Evidence for early interactive processing. *Psychological Science, 14,* 328–333.

Simpson, G. B., & Krueger, M. (1991). Selective access of homograph meanings in sentence context. *Journal of Memory and Language, 30,* 627–643.

Strand, E.A. & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In Gibbon, D. (ed), *Natural Language Processing & Speech Technology: Results of the 3rd Konvens Conference, Bielefield, Oct 1996*. Walter de Gruyter, Berlin.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially-weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology, 4,* Article 1015. doi: 10.3389/fpsyg.2013.01015

Tabossi, P. (1988). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language, 27,* 324–340.

Tesink, C. M. J. Y., Magnus Petersson, K., van Berkum, J. A., van den Brink, D., Buitelaar, J. K. & Hagoort, P. (2009). Unification of Speaker and Meaning in Language Comprehension: An fMRI Study. *Journal of Cognitive Neuroscience, 21,* 2085-2099.

van den Brink, D., van Berkum, J. J. A., Bastiaansen, M. C. M., Tesink, C. M. J. Y., Kos, M., Buitelaar, J. K., & Hagoort, P. 2012. Empathy matters: ERP evidence for inter-individual differences in social language processing *Social Cognitive and Affective Neuroscience, 7,* 173-183. doi: 10.1093/scan/nsq094

Van Berkum, J. J. A.; Brown, C. M.; Zwitserlood, P.; Kooijman, V.; Hagoort, P.(2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, Vol 31,* 443-467. doi:10.1037/0278-7393.31.3.443

Van Berkum, J. J. A., Brink, D. van den, Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20,* 580-591.

Vu, H., Kellas, G., Petersen, E., & Metcalf, K. (2003). Situation-evoking stimuli, domain of reference, and the incremental interpretation of lexical ambiguity. *Memory & Cognition, 31,* 1302–1315.

Walker, A. & Hay, J. (2011). Congruence between ‘word age’ and ‘voice age’ facilitates lexical access. *Laboratory Phonology, 2*, 219-237.

Table 1. Mean ratings of voice age and gender for the four speakers used in Experiments 1 – 3.

	Masc	Fem	Adult	Childlike
Male speaker	**6.82**	1.07	6.36	1.22
Female speaker	1.38	**5.88**	4.88	2.14
Adult speaker	1.81	6.06	**6.40**	1.39
Child speaker	3.70	3.49	1.26	**6.27**

Note: Bold-faced ratings indicate the intended proto-typical social category for each speaker. Raters used a 7-point scale, with 7 indicating a high level of each attribute.

T*able 2. Summary of effects from latency model for Experiments 1 - 3.*

Experiment		Df	Sum Sq	Mean Sq	F value
Experiment 1					
	Congruency	1	526,161	526,161	4.63*
	Bias	3	9,066	3,022	0.027
	Congruency*Bias	3	1,742,069	580,690	5.11*
Experiment 2					
	Congruency	1	43,938	43,938	0.76
	Bias	3	143,940	47,980	0.83
	Congruency*Bias	3	76,606	25,535	0.44
Experiment 3					
	Congruency	1	134,975	134,975	1.56
	Bias	3	52,357	17,452	0.20
	Congruency*Bias	3	542,106	180,702	2.09

Note: Significant F values are indicated by an asterisk.

Table 3. *Growth curve analysis for fixations to target pictures in Experiment 1.*

	Estimate	Std..Error	T value	p
Intercept	0.1772	0.0115	15.35*	<<.001
ot1	0.3551	0.0321	11.07*	<.001
ot2	0.1126	0.0177	6.35*	<.001
ot3	-0.0045	0.0133	-0.34	0.736
Condition nc	-0.0326	0.0134	-2.42*	0.015
ot1:condnc	-0.0261	0.0422	-0.62	0.536
ot2:condnc	0.01097	0.0243	0.45	0.652
ot3:condnc	0.00834	0.0148	0.56	0.573

Table 4. *Growth curve analysis for fixations to target pictures in Experiment 2.*

	Estimate	Std..Error	T value	p
Intercept	0.45496	0.0127	35.9*	< .01
ot1	0.6503	0.0513	12.68*	<.01
ot2	-0.4249	0.0426	-9.97*	<.01
ot3	-0.2055	0.0234	-8.80*	<.01
Condition nc	-0.0079	0.0144	-0.55	0.58
ot1:condnc	-0.08198	0.0529	-1.55	0.12
ot2:condnc	-0.0058	0.0517	-0.11	0.91
ot3:condnc	0.0058	0.0225	0.26	0.80

Table 5. *Growth curve analysis for fixations to target pictures in Experiment 3.*

	Estimate	Std..Error	T value	p
Intercept	0.3683	0.0123	29.90	<.01
ot1	0.5163	0.0511	10.11	<.01
ot2	-0.2855	0.0391	-7.29	<.01
ot3	-0.1749	0.0202	-8.64	<.01
Condition nc	0.0105	0.0144	0.73	0.015
ot1:condnc	-0.0156	0.0482	-0.32	0.536
ot2:condnc	-0.0355	0.0410	-0.86	0.652
ot3:condnc	-0.01997	0.0203	-0.98	0.573

Table 6. Summary of effects from latency model combining Experiments 2 and 3.

	Df	Sum Sq	Mean Sq	F value
Congruency	1	18,408	18,408	0.25
Noise	1	337,153	337,153	4.54*
Bias	3	131,312	43,771	0.59
Congruency* Noise	1	158,280	158,280	2.13
Noise*Bias	3	317,012	105,671	1.42
Congruency*Bias	3	320,438	106,813	1.44
Congruency*Noise*Bias	3	274,230	91,410	1.23

Figure 1. Fixations to the target picture, phonological competitor, and filler pictures in Experiment 1 for the congruent condition (Fig. 1A) and the incongruent condition (Fig 1B). The line corresponding to "fillers" is averaged over the two filler pictures. For this reason, and because participants were often still fixated on the center of the screen (early in the time window) or had already looked at the target to complete the trial (late in the time window), the sum of the fixations at a given time point do not sum to 1.

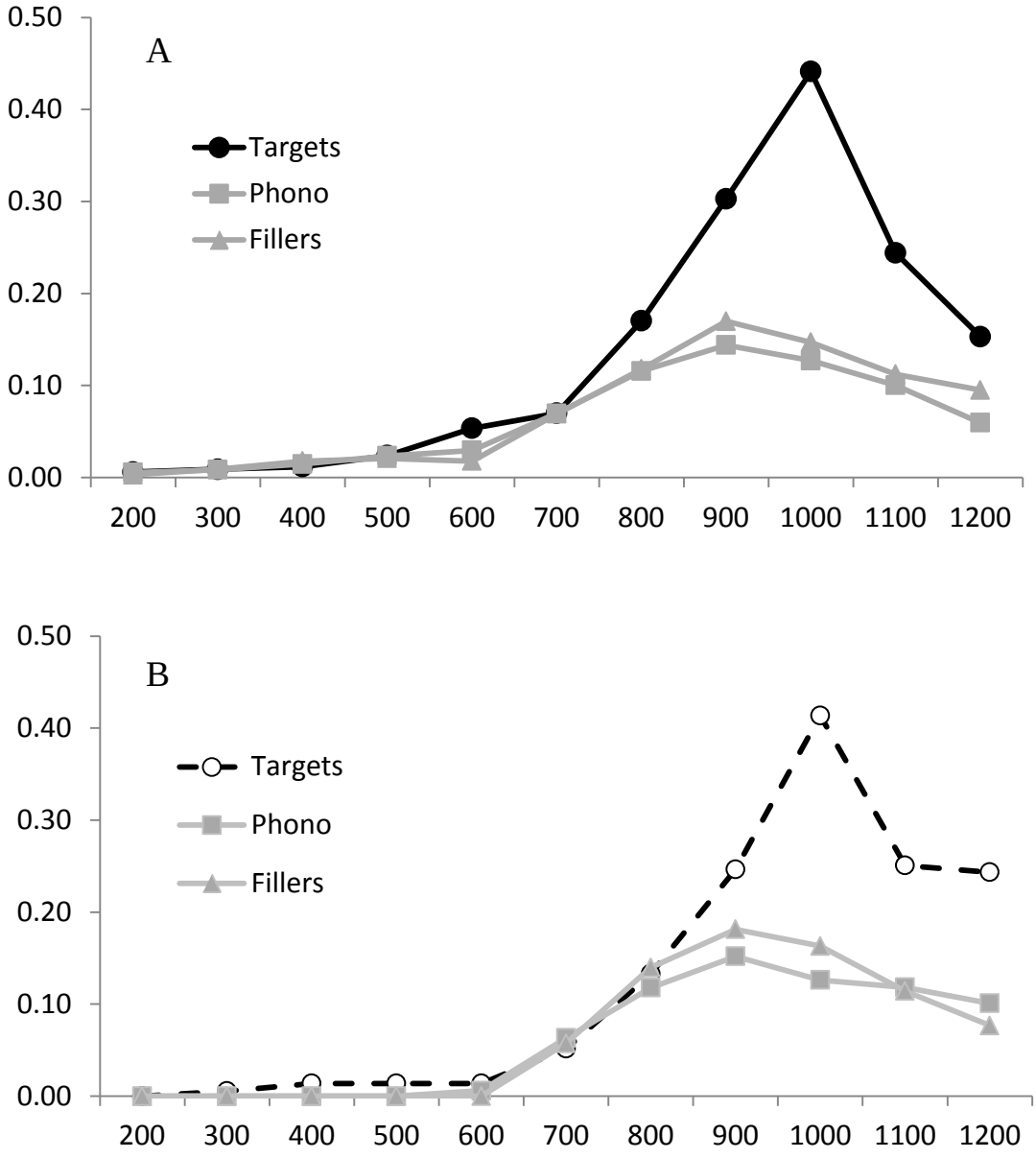


Figure 2. A visual summary of the growth curve analysis for fixations on the target images during the critical interval of Experiment 1. The points represent the observed data (with standard errors). The lines represent the fitted model.

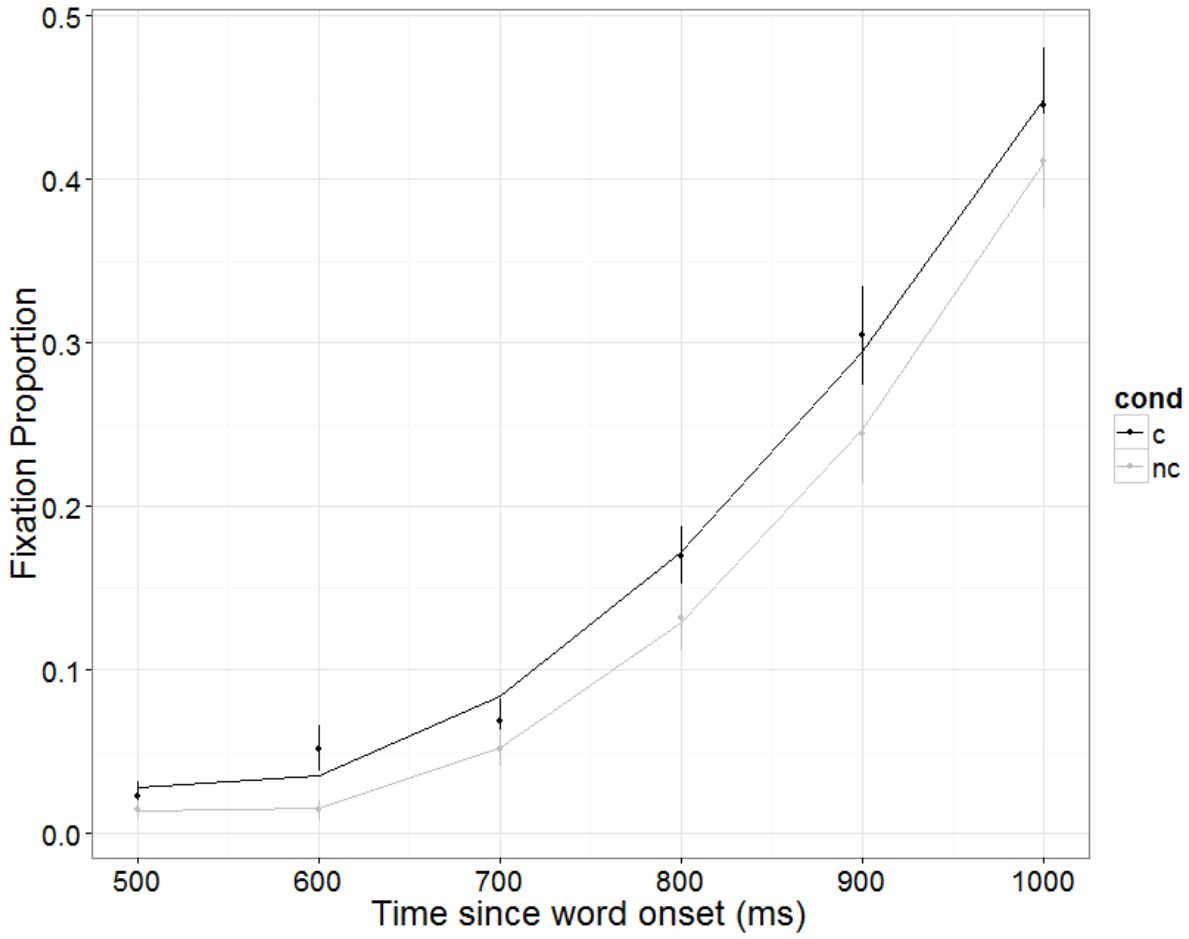


Figure 3. Fixations to target picture in congruent and incongruent conditions in Experiment 2, 200 to 1200 ms after homophone onset. The points represent observed data and the lines represent the fitted model.

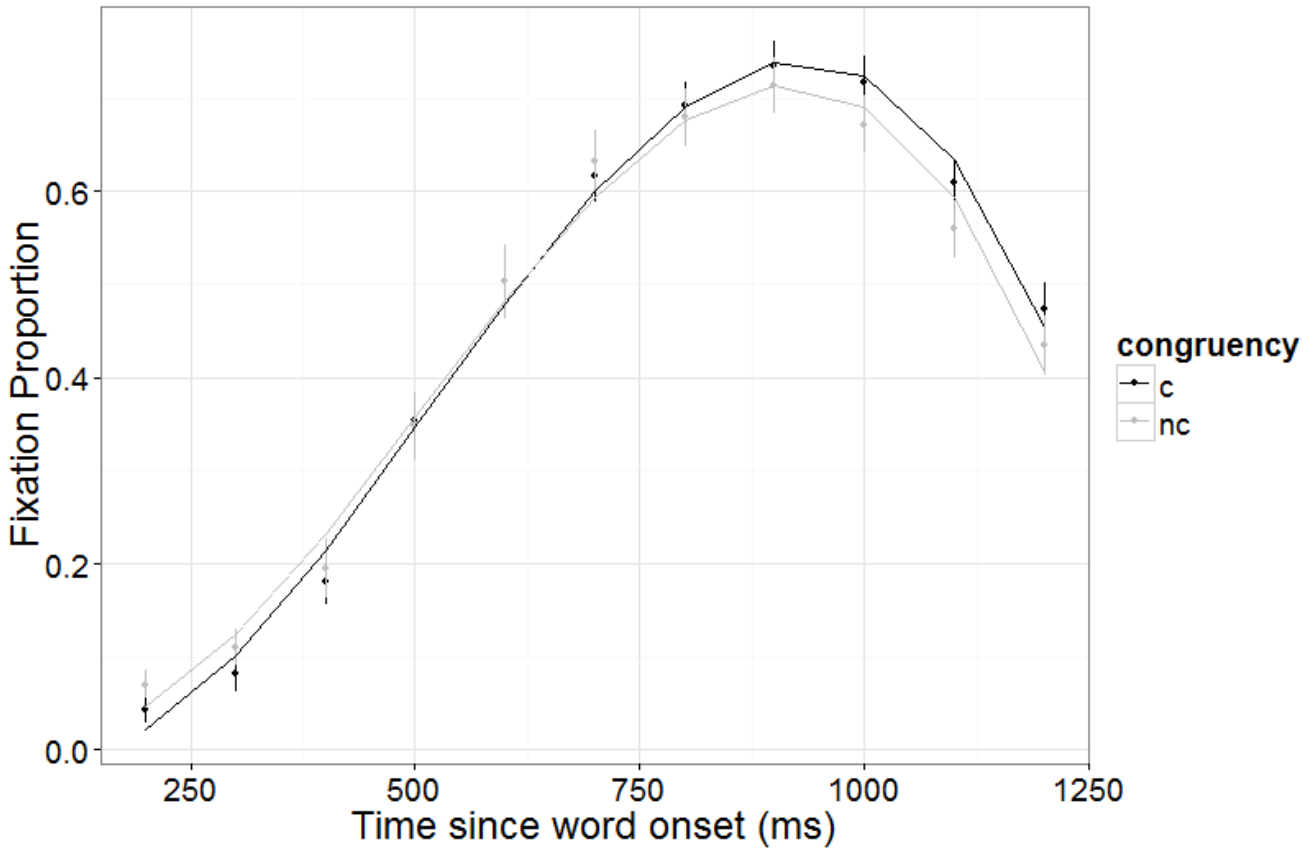


Figure 4. Fixations to target picture, phonological competitor, and filler pictures in Experiment 2 for the congruent condition (Fig. 4A) and the incongruent condition (Fig 4B).

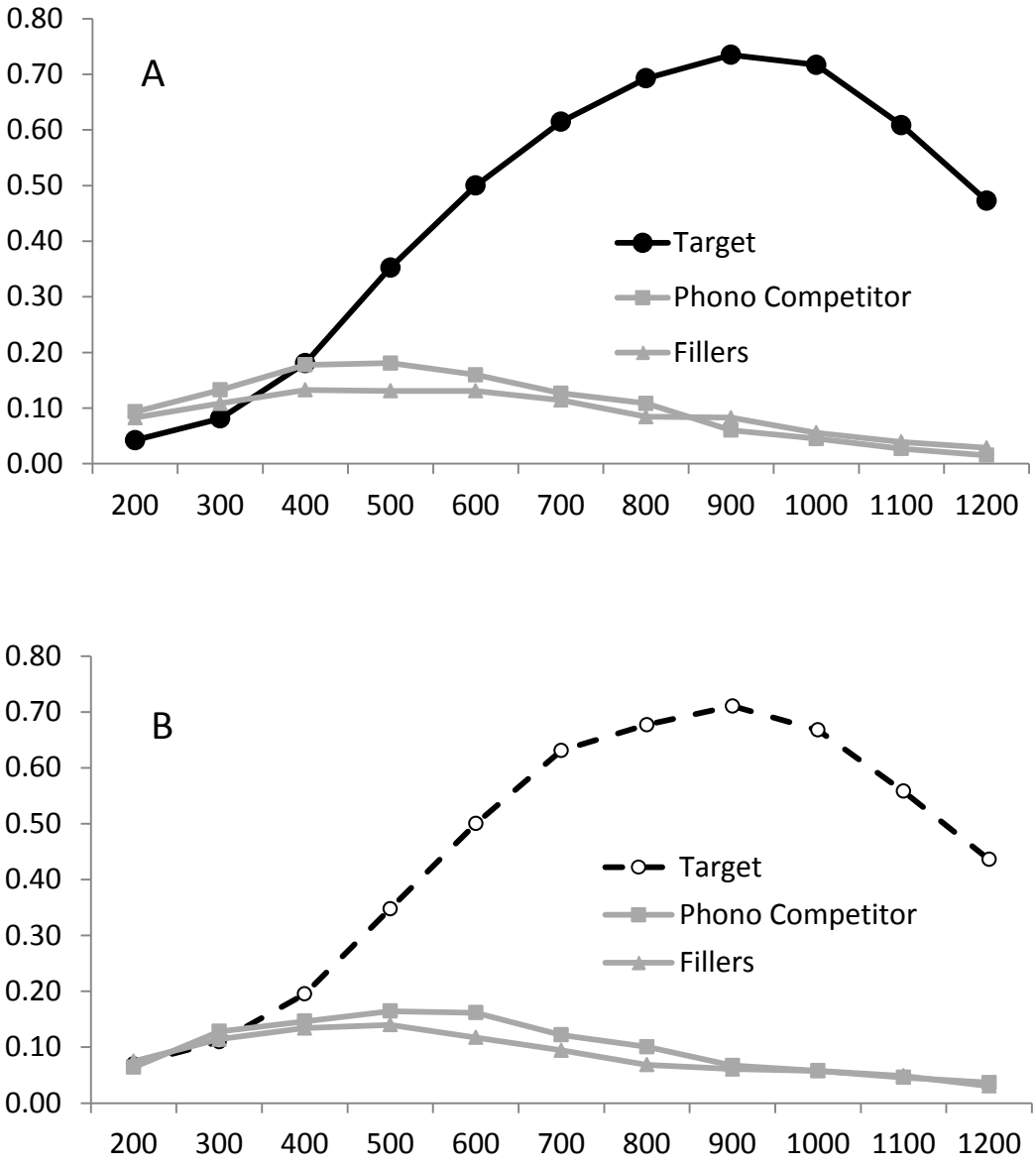


Figure 5. Looks to the target image in the congruent and incongruent conditions. The plot points represent the means of the observed data; the lines represent the fitted model.

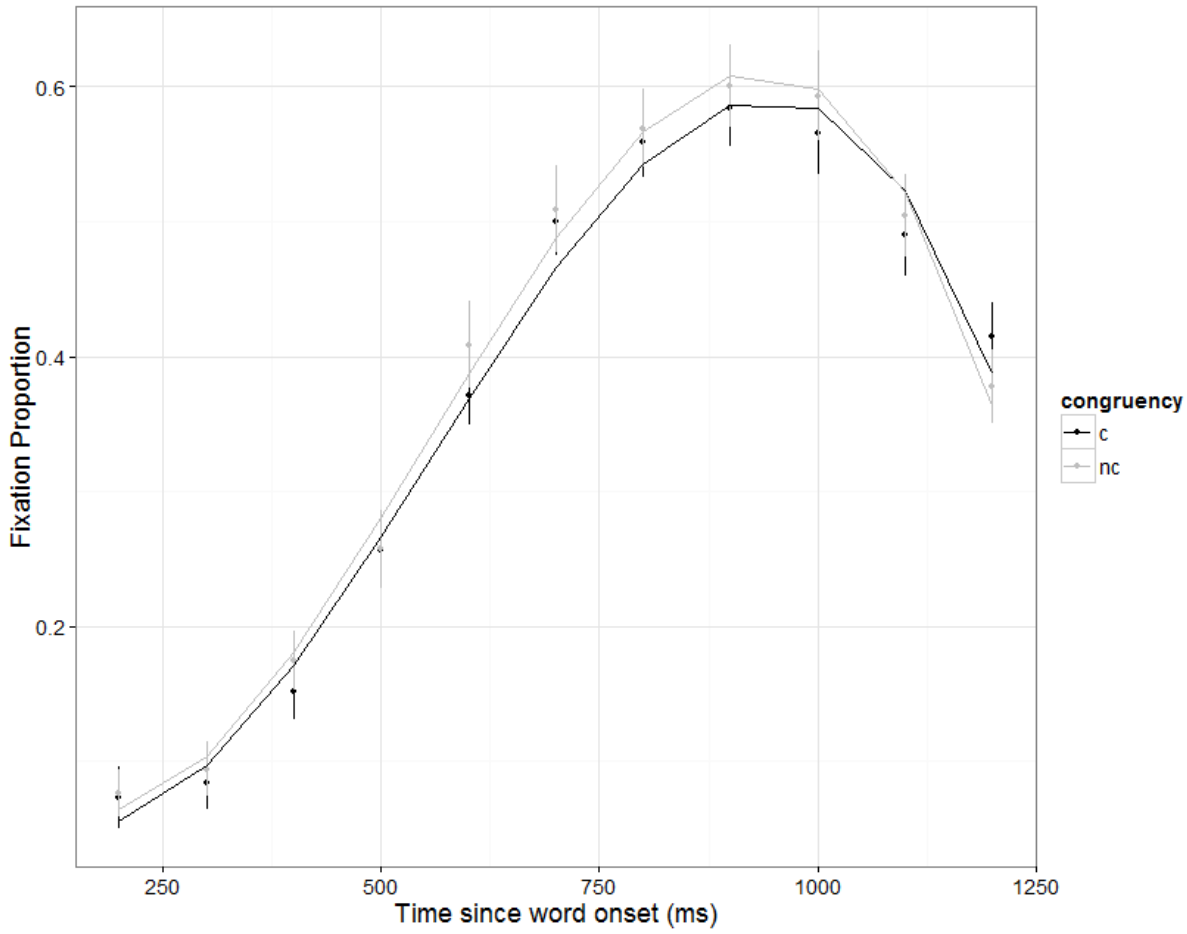
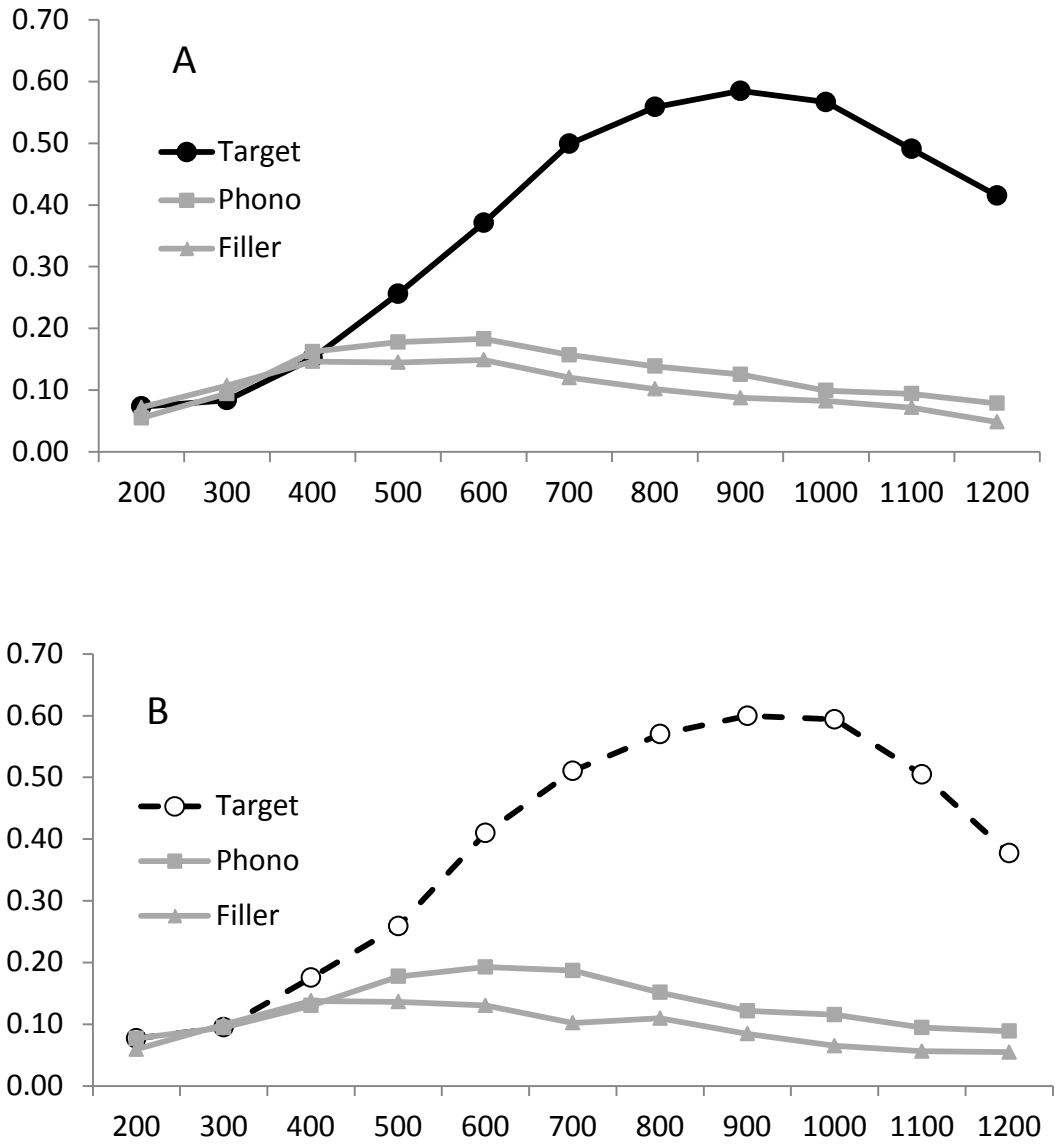


Figure 6. Fixation patterns for all 3 image types in Experiment 3, in the congruent voice (Fig 6a) and incongruent voice (Fig 6b) conditions.



Appendix

The linguistic stimuli for Experiment 1 are listed below, with the attested social bias of the homophone sense. The homophones were always the final word in the sentence, and are the same homophones used in Experiments 2 and 3.

At the end of the month, I write out a check.	adult
When I feel creative, I improvise on my sax.	adult
They played on the floor with the jacks.	child
She wanted to play so she spun the top.	child
At the gathering there was a table for her presents.	child
On a hot summer day, I like a sundae.	child
My favorite character in the story was the fairy.	child
In the morning, I usually eat cereal.	child
During the afternoon, I like to go to the park.	child
When it was my turn, I swung the bat.	child
I will attach the sign with my nails.	male
I'm going to the woods with this bow.	male
In the open field, they saw a deer.	male

They will lift the car with the jacks.	male
For an important meeting, you should wear a nice tie.	male
When I'm really hungry, I like a steak.	male
As they approached the water, they saw a big sail.	male
To warm up, I swung my 9-iron over the tee.	male
I will grab the splinter with my nails.	female
I'm going to wrap the box with this bow.	female
When I go swimming, I take off my ring.	female
When you dress up, you should wear your hose.	female
Part of good grooming is having clean hair.	female
As they walked on Main Street, they saw a big sale.	female
On a fancy date, don't forget some make-up.	female