Cognition 172 (2018) 134–143

Contents lists available at ScienceDirect

Cognition

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

Original Articles

# Understanding environmental sounds in sentence context

Sophia Uddin*, Shannon L.M. Heald, Stephen C. Van Hedger, Serena Klos, Howard C. Nusbaum

*Department of Psychology, The University of Chicago, 5848 S. University Ave., Chicago, IL 60637, USA*

ARTICLE INFO

*Keywords:*
Constraint
Language
Recognition
Context
Speech perception
Environmental sound perception

ABSTRACT

There is debate about how individuals use context to successfully predict and recognize words. One view argues that context supports neural predictions that make use of the speech motor system, whereas other views argue for a sensory or conceptual level of prediction. While environmental sounds can convey clear referential meaning, they are not linguistic signals, and are thus neither produced with the vocal tract nor typically encountered in sentence context. We compared the effect of spoken sentence context on recognition and comprehension of spoken words versus nonspeech, environmental sounds. In Experiment 1, sentence context decreased the amount of signal needed for recognition of spoken words and environmental sounds in similar fashion. In Experiment 2, listeners judged sentence meaning in both high and low contextually constraining sentence frames, when the final word was present or replaced with a matching environmental sound. Results showed that sentence constraint affected decision time similarly for speech and nonspeech, such that high constraint sentences (i.e., frame plus completion) were processed faster than low constraint sentences for speech and nonspeech. Linguistic context facilitates the recognition and understanding of nonspeech sounds in much the same way as for spoken words. This argues against a simple form of a speech-motor explanation of predictive coding in spoken language understanding, and suggests support for conceptual-level predictions.

## 1. Introduction

One of the hallmarks of both spoken and written language is the interaction of word recognition with the meaning of linguistic context (Morris & Harris, 2002; Simpson, Peterson, Casteel, & Burgess, 1989). A long-known example is semantic priming, in which words are recognized faster when preceded by a related word rather than an unrelated word (Hutchison et al., 2013; Meyer & Schvaneveldt, 1971). Meaningful sentence context affects word recognition as well. Gating studies, in which a spoken word is presented incrementally in small sound segments of increasing length, have shown that in a highly constraining sentence context (as opposed to a vague context), people need to hear less signal to identify a spoken word (Grosjean, 1980; Tyler & Marslen-Wilson, 1986). Additionally, when people are asked to complete a sentence ending, they supply a word faster for a highly constrained sentence context than for a low constraint context (Staub, Grant, Astheimer, & Cohen, 2015).

Why is word recognition influenced by linguistic context? Extant word recognition models incorporate the effects of context information on lexical knowledge to varying degrees (see Dahan & Magnuson, 2006 for a review). Some models suggest that bottom-up input (e.g., the acoustic waveform of a spoken word or the visual input of a printed word) is the primary determining factor in the recognition process (e.g., Norris, 1994; Norris & McQueen, 2008). In these models, input is processed in a feed-forward manner through a series of transformations until a word is recognized, and it is only at late stages, when the recognized word's meaning is being assessed, that it is integrated with and constrained by its surrounding context. Some models draw on evidence from priming studies to argue for a two-stage process in which bottom-up input causes widespread activation of many candidate words that could be consistent with the input, but are not constrained to be consistent with the broader context (for example, the word "bug" primes both "ant" and "spy," even if the context suggests only the first interpretation; Swinney, 1979). According to such models, context then acts later, in the second stage of the model or "selection phase", by facilitating the process of narrowing down from the population of activated candidates to the word that best fits the context (Swinney, 1979; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

In contrast to these "input driven" models, interactive recognition models allow for continuous, on-line effects of context on word recognition. In such models, higher-level information, such as semantic associations, can alter processing at lower levels in a top-down manner via continuous integration (e.g., McClelland & Elman, 1986; Mirman, McClelland, & Holt, 2006) Shillcock and Bard (1993) were early critics of the modular, two-stage account, arguing that for closed-class words, immediate (as opposed to delayed) context effects support a continuous

* Corresponding author.
*E-mail address:* sophiauddin@uchicago.edu (S. Uddin).

https://doi.org/10.1016/j.cognition.2017.12.009
Received 26 September 2016; Received in revised form 16 November 2017; Accepted 14 December 2017

integration model. Further, eye tracking and fMRI studies have found context effects extremely early in processing, before other models incorporate context effects, and even before the bottom-up input unambiguously identifies a single word (Dahan, Magnuson, & Tanenhaus, 2001; Dahan & Tanenhaus, 2004; Magnuson, Tanenhaus, & Aslin, 2008; Revill, Aslin, Tanenhaus, & Bavelier, 2008). These studies suggest that lexical representations and semantic associations are being accessed simultaneously and integrated with each other continuously.

In recent years, interactive recognition models have been reinterpreted in light of predictive coding. In predictive coding accounts, language comprehension rests on neural predictions, based on context or prior knowledge, that are continuously compared against input as it is being processed (e.g., Bonhage, Mueller, Friederici, & Fiebach, 2015; DeLong, Urbach, & Kutas, 2005; McRae, Hare, Elman, & Ferretti, 2005; Metusalem et al., 2012; Pickering & Garrod, 2007). While some (e.g., Pickering & Garrod, 2007) argue that the speech motor system is integral to predictive coding, this view is by no means universal (see Hickok, 2012). ERP data from Federmeier and Kutas (1999) suggests that context allows the prediction of semantic features for upcoming words. However, it is possible that linguistic predictions could instead be happening at the level of sensory (e.g., auditory or visual) representations (cf Lewis & Bastiaansen, 2015). It is also possible that predictions involve both semantic and sensory information (Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Kuperberg & Jaeger, 2016; Lupyan & Clark, 2015; McRae et al., 2005).

The notion that semantic associations influence ongoing and subsequent lexical processing (and vice versa) is supported by a substantial body of work on cross-modal effects. One example of this sort of cross-modal interaction is analog acoustic expression, the phenomenon in which modulations in pitch and speaking rate in speech affects the listener's understanding of the message (e.g. Shintel, Nusbaum, & Okrent, 2006). Information from musical underscoring can affect speech understanding in a similar manner (Hedger, Nusbaum, & Hoeckner, 2013). The effects of non-linguistic information on linguistic interpretation are not confined to the auditory modality. Tanenhaus and colleagues have used eye tracking to demonstrate that listeners make rapid on-line use of visual scene context in order to disambiguate spoken verbal instructions (Chambers, Tanenhaus, & Magnuson, 2004; Tanenhaus et al., 1995). Such cross-modal effects on language have also been demonstrated via priming studies, in which visual or spoken words can facilitate processing of environmental sounds and vice versa (Frey, Aramaki, & Besson, 2014; Orgs, Lange, Dombrowski, & Heil, 2006, 2007; van Petten & Rheinfelder, 1995). Both words and environmental sounds have also been found to prime recognition of pictures (Chen & Spence, 2011; Schneider, Engel, & Debener, 2008). Concepts associated with words can also influence processing in other domains, as when words describing a particular direction of motion (such as the word "approach") affect visual motion perception (Meteyard, Bahrami, & Vigliocco, 2007). Even when concepts are conveyed in a complex, non-linguistic way (e.g., an auditory scene), they can bias the people's verbal labels for ambiguous environmental sounds (Ballas & Mullins, 1991). Thus, there is strong evidence that such cross-modal interactions occur bidirectionally, such that non-linguistic contextual information can cross-modally facilitate spoken word processing, and verbal context can facilitate processing of non-linguistic stimuli.

Despite the extensive documentation of cross-modal interactions between non-linguistic and linguistic information, the mechanisms behind these effects remain unclear. One possibility is that participants are covertly naming non-linguistic stimuli in order to guide processing words. This possibility is favored by a modular account of language processing, as according to this viewpoint, non-linguistic information cannot interact with encapsulated language modules until it is translated into linguistic information. It seems unlikely, however, that this is the case Potter, Kroll, Yachzel, Carpenter, and Sherman (1986) asked whether printed sentences containing a picture substituted for a noun affected the speed and accuracy of plausibility judgments about the sentences. They reasoned that if pictures directly access the same system of concepts as words, rather than first being covertly named, then response times for plausibility judgments should be similar for "rebus" sentences (those containing a picture substituting for a word) and all-word sentences. This was indeed what they found. The results could not be easily attributed to covert naming, as previous work has demonstrated that picture naming takes too long to be occurring in Potter's paradigm (cf Oldfield & Wingfield, 1965). Other work also suggests against covert naming as the mechanism for context effects. The studies by Chambers et al. (2004) and Tanenhaus et al. (1995) rely on sufficiently complex visual scenes that covert naming alone would not resolve the ambiguities present. Finally, it is highly unlikely that covert naming could explain analog acoustic expression effects, as listeners would have to translate the metaphoric meaning present in vocal pitch or rate information directly into words.

If covert naming is not responsible for the cross-modal priming effects that have been previously described, how does this process work? It is possible that, as suggested by Potter and colleagues, words and non-verbal stimuli such as pictures access a single conceptual system that is not grounded in language. In other words, the same neural representations of semantic information could be accessible via words and other meaningful non-verbal stimuli. Work by Zwaan and colleagues describes an effect opposite to covert naming, in which words activate "mental pictures" of the objects to which they refer (Zwaan, Stanfield, & Yaxley, 2002), providing support for Potter's hypothesis that words draw on a general conceptual system that is also used by nonverbal stimuli. In terms of a predictive coding framework, this would mean that predictions are sufficiently amodal (or multimodal) to interact easily with information from different domains. It is important to note that many models of word recognition are largely concerned with information involving phonemes and lexical representations, and have not been extended to representations that involve general concepts or "mental pictures" (Mcclelland, Mirman, & Holt, 2006; Mirman et al., 2006; Norris & McQueen, 2008; Strauss, Harris, & Magnuson, 2007) although it is certainly possible to do so, especially considering the aforementioned studies, which suggest that this non-lexical information is readily and perhaps obligatorily activated by words.

In the present experiment, we asked how recognition of recognizable and meaningful, but non-linguistic, environmental sounds would be affected by linguistic context by using spoken sentence frames that were completed as a sentence by either a spoken word or an environmental sound. An account of language understanding that isolates speech processing as a separate system from a broader conceptual system predicts that integrating non-linguistic inputs with preceding sentence context should be more difficult than integrating spoken word inputs. Non-linguistic information should be integrated as post-perceptual problem solving, requiring a covert naming step. This might incur heavy processing costs (over 500 ms for covert naming, cf Oldfield & Wingfield, 1965). Based on prior research, however, it seems unlikely that strictly isolated speech processing would occur. Not only have cross-modal effects involving rapid interaction of many types of non-linguistic information with language been documented, but recent research has suggested that words and meaningful non-linguistic stimuli may have more in common in processing than previously thought given the neural resources involved in understanding both (Cummings et al., 2006; Dick, Krishnan, Leech, & Saygin, 2016; Leech & Saygin, 2011; Saygin, 2003; Saygin, Dick, & Bates, 2005). However, there is little research on how environmental sounds are understood, especially in comparison to speech sounds, and few studies directly comparing recognition and understanding of these two classes of sounds under a common contextual constraint.

Using this paradigm, we can measure whether the recognition or understanding of an environmental sound in a sentence frame relies on a reallocation of attention beyond what might be found for re-orienting to a new talker. Recognizing speech when there is a change in the talker in a sequence of utterances increases recognition time by about 40 ms (Heald & Nusbaum, 2014b; Nusbaum & Morin, 1992; Wong, Nusbaum, & Small, 2004). Moreover, a similar recognition cost around 40 ms is found for recognizing musical notes when there is a change in instrument (Van Hedger, Heald, & Nusbaum, 2015). Given this recognition cost for shifts in signal source (voice or timbre) we might predict a similar cost or greater for recognition of an environmental sound in the context of speech.

Further, how does sentential (linguistic) constraint affect processing of environmental sounds roughly matched in meaning to spoken words? Spoken words following semantically constraining sentence frames will be recognized with less signal and responded to faster than words following less constraining frames, but whether this "constraint benefit" will occur to the same degree for environmental sounds is an open question. If the constraint operates at a purely linguistic level, we would not expect to see similar effects for environmental sounds.

Our experiments also allow us to address questions about how constraint may be acting via predictive coding. There is evidence that in particular situations, (e.g., when spoken word forms are predicted by context, or a non-linguistic motor action is predicted), the speech motor system generates predictions that aid comprehension (see Pickering & Garrod, 2007, 2013). However, the substitution of environmental sounds for spoken words presents a different kind of situation: the recognition and comprehension of the nonspeech sound pattern cannot be aided directly by the speech motor system, because the nonspeech sound patterns are not vocally generated. Thus, the question would appear to fall more in the realm of predicting an appropriate non-linguistic motor action based on the linguistic form combined with the environmental sound. However, participants in the current study are not deciding how to act in this situation, but simply recognizing the nonspeech sound or understanding the sentence frame-plus-sound. Given that this falls outside the explanatory domain of speech-motor system predictive coding, parallel results for increasing sentence constraint in recognition or comprehension for spoken words and meaningful nonspeech sounds would pose a theoretical challenge.

Similarly, we can also use this experiment to ask, outside of a speech-motor framework, how predictive coding operates to constrain processing. The specific acoustics of environmental sounds are likely much harder to predict than the acoustics of the same speaker from the sentence frame saying the final word of the sentence, simply because the range of possible environmental sounds that could fit the intended meaning is larger. Thus, it follows that if neural predictions derived from context are largely at the level of sensory representations, constraining sentence contexts should be more helpful to spoken words than environmental sounds. However, if such predictions are more conceptual in nature, we would expect similar constraint benefits for spoken words and environmental sounds.

## 2. Experiment 1

The gating paradigm (e.g. Grosjean, 1980) has been used to measure how much of a spoken word waveform is needed for recognition. In general, gating studies show that listeners can correctly identify spoken words even before the end of the word is heard, and that less of the word is needed for recognition in sentence context than in isolation (Grosjean, 1980). The present study used the gating paradigm in a similar way: to compare the effects of sentence context in facilitating identification for spoken words and non-linguistic (environmental) sounds. The main question for this experiment is whether the benefit of sentential constraint is similar for environmental sounds and spoken words.

### 2.1. Methods

#### 2.1.1. Participants

There were 131 participants (78 female). Participants were from the University of Chicago community (mean: 19.89 years, range: 18–27 years); this includes students, staff, and residents of the surrounding area. Informed consent was obtained in accordance with IRB-approved protocol, and participants were compensated with their choice of one course credit or $10 per hour of their time. Due to the linguistic and auditory nature of the task, participants were limited to those who reported speaking English as a first language, and who reported having normal hearing.

#### 2.1.2. Stimuli

Stimuli consisted of sentence frames recorded at 44.1 kHz by an adult male native speaker of American English. There were two levels of sentence frame constraint, such that half were relatively constraining of the final word (high cloze probability, median = 0.87, IQR = 0.25) and half were less constraining (low cloze probability, median = 0.16, IQR = 0.33). We will refer to these two categories as "specific" (high cloze probability) and "general" (low cloze probability). Cloze probability was determined based on sentence completions from 66 Amazon Mechanical Turk participants. The last word of each sentence was the *target*, and was recorded separately from the sentence frame to avoid co-articulation confounds. There were 32 targets.

Each target word was a noun that could be represented by an environmental sound (e.g. "sheep", and the sound of a sheep bleating, Appendix Table A1). Corresponding environmental sounds were taken from online databases such as soundbible, and if necessary resampled to 44.1 kHz. All sounds were then normalized to the same RMS level as the sentence frames and target words using Matlab. A small survey of lab members was conducted to ensure that the sounds were identifiable when heard in isolation. Mean duration of targets was 0.502 s for spoken words and 0.838 s for environmental sounds (see Supplement). Eight of the 32 environmental sounds involved repetition (e.g., the sound of a siren involves repeating pitch oscillations; see Supplement). Sentence frames and target sounds were digitally spliced together to create complete sentences. Half the resulting sentences terminated in spoken word targets, and half terminated in categorically matched environmental sounds. Stimuli were presented at 65–70 dB over stereo headphones.

#### 2.1.3. Task

The task was based on Grosjean (1980); participants heard the targets in progressively increasing 20 ms waveform increments, and the task was to identify the target via a freely typed identification response.

There were six groups of participants tested using a 2 × 3 design crossing target type (between-subjects: sound or word) and context (between-subjects: general, specific, or isolated – i.e. no sentence context—target). There were between 20 and 23 participants in each group (Table 1). Participants heard each target once for a total of 32 trials; the order of these trials was randomized. In the general and specific groups, participants heard the targets after the appropriately constrained sentence frame. In the isolated groups, they heard only the targets.

Targets, either isolated or at the ends of sentence frames (not sentence frames themselves) were presented in successively increasing segment lengths by 20 ms increments until either (1) the entire sound

**Table 1**
Numbers of participants in each group for experiment 1 after the exclusion of the three participants as described in *Data Analysis* for Experiment 1.

| Condition | n |
|---|---|
| Isolated Sounds | 23 |
| General Sounds | 22 |
| Specific Sounds | 20 |
| Isolated Words | 21 |
| General Words | 22 |
| Specific Words | 20 |

was presented, or (2) the participant's identification responses remained stable for 20 gates in a row. As soon as either (1) or (2) was reached, the participant heard the whole sound, and was asked to identify it one last time.

### 2.1.4. Data collection

Participants' responses and corresponding gates were collected in Matlab 2014 with Psychtoolbox 3 (Brainard, 1997; Kleiner et al., 2007) and recognition accuracy was scored by hand. For each target, the gate of recognition was defined as the gate of the earliest correct response, to conservatively determine the minimal signal supporting recognition. For each target, recognition points were calculated by converting gates of recognition to seconds and averaging across subjects in the same condition. This resulted in three recognition points for each target: general context, specific context, and isolated target.

### 2.1.5. Data analysis

Three participants were excluded: one for reporting not being a native English speaker after the experiment (condition: environmental sounds/specific frames), one for previous participation in the study (condition: spoken words/general frames) and one for not following instructions (condition: environmental sounds/specific frames).

Percent correct responses for each target in each condition were calculated across participants. While we expected that most participants would be close to ceiling for words, we expected some misidentification of environmental sounds, particularly in the isolated target condition. Any target that dropped below 70% correct in any condition across participants was removed from further analysis. Five environmental sounds (baby laughing, creaky door, clock ticking, papers ruffling, and sword being unsheathed) and one word ("horn") were excluded from the final analysis due to this level of poor recognition in the isolated condition. In conditions with sentence context, recognition performance was always well above this level.

After exclusions of low-accuracy targets and problem participants, the mean recognition point for each target in each context was calculated across subjects, and data points outside 2.5 standard deviations from the mean were removed. We excluded recognition points from the specific context condition from further analysis. For a majority of the trials with specific sentence frames, participants answered correctly on the first gate (after hearing only 20 ms of the target), regardless of target type (sound or word). For highly constrained frames, guessing the ending may be too easy when there are no foils or distractors, leading to a ceiling effect. Thus, further examination of the data from the specific sentence frames is not informative.

A Repeated Measures ANOVA was performed on the resulting recognition gate data. The factors modeled included constraint level (isolated versus general frame), and target type (word or environmental sound). Excluding words corresponding to the five poorly recognized environmental sounds, and excluding the environmental sound corresponding to the excluded word "horn" did not substantially change the ANOVA results, so these stimuli are included in further analyses.

## 2.2. Results

There was strong evidence for a context effect on recognition points for both environmental sounds and words (Fig. 1), such that for both target types, adding sentence context significantly decreased time to recognition. Participants recognized targets in isolation after hearing an average of 272 ms of waveform; this dropped to 139 ms of waveform for targets occurring after a general sentence frame. Thus context significantly reduces the amount of waveform needed to recognize acoustic targets ($F\ (1,\ 26) = 115.1,\ p\ <\ .001,\ d = 4.2$). For spoken words, sentence context reduced the amount of waveform needed for recognition from 261 ms to 141 ms. This reduction of 120 ms by general sentence context compared to isolated words is similar to the effects of general context compared to isolated targets reported in previous gating studies (Cotton & Grosjean, 1984; Grosjean, 1980; Tyler & Wessels, 1983).

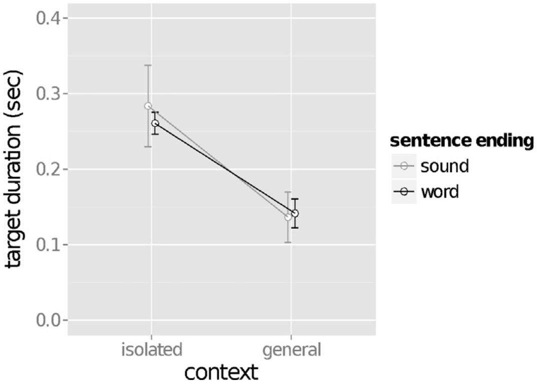

**Fig. 1.** Mean duration of target needed for recognition. Sentence context shortens the amount of target waveform needed for participants to recognize both environmental sounds and words. Error bars represent ± 1 standard error of the mean (SEM).

It is also the case that general sentence context reduces the amount of waveform needed for the recognition of identifiable environmental sounds by 148 ms: from 284 ms for isolated sounds to 136 ms for sounds in linguistic context. Clearly, the information in sentence context is informative about the identity of meaningful sounds.

While recognition points for sounds in isolation appear to be slightly later than those for words, and slightly earlier in the presence of a general sentence context (Fig. 1), there was no significant difference between recognition points for sounds and words ($F\ (1,\ 26) = 0.007$, $p\ >\ .25,\ d = 0$). Moreover, there was no significant interaction between context and target type (Fig. 1, $F\ (1,\ 26) = 0.64,\ p\ >\ .25$, $d = 0.31$).

Our general sentence frames represented a range of cloze probabilities. If the reduction in recognition point observed for targets in sentence context is truly due to facilitation involving the conceptual meaning of the sentence (as opposed, for example, to a purely psychophysical effect of any sound preceding the target), there should be an inverse relationship between cloze probability and recognition point for the sounds and words presented in sentence context. This was indeed the case; there were significant nonlinear inverse relationships as shown by Spearman's rank order correlations (Fig. 2, Sounds: $\rho\ [25] = -0.45,\ p = .02$; Words: $\rho\ [29] = -0.71,\ p = 7.3\text{e}-6$). The strength of these inverse relationships was not significantly different between sounds and words ($z = 1.45,\ p = .15$).

## 2.3. Discussion

Sentence context significantly reduces the amount of signal needed for recognition of both spoken words and environmental sounds, even when the sentence frame is general. This reduction happens to the same extent for both sounds and words; there is no evidence for any interaction between context and target type. From a predictive coding perspective, these results suggest that sentence frames support neural predictions that are mainly conceptual rather than sensory, because context helps word and sound recognition to the same extent, even though sensory predictions should be less precise for environmental sounds. In the general constraint condition, the inverse relationship between recognition point and cloze probability indicates that listeners can use fine-grained meaning constraint information to facilitate processing of both sounds and words to a similar extent. This relationship also indicates that the earlier recognition points to items in general context (as opposed to isolated items) is not due merely to sound preceding the targets, but to the way sentence constraint and meaning interact with the interpretation of the target sound or word.

The gating paradigm is intended to mimic the recognition of a sound as it unfolds in time. However, sounds do not stop in midstream

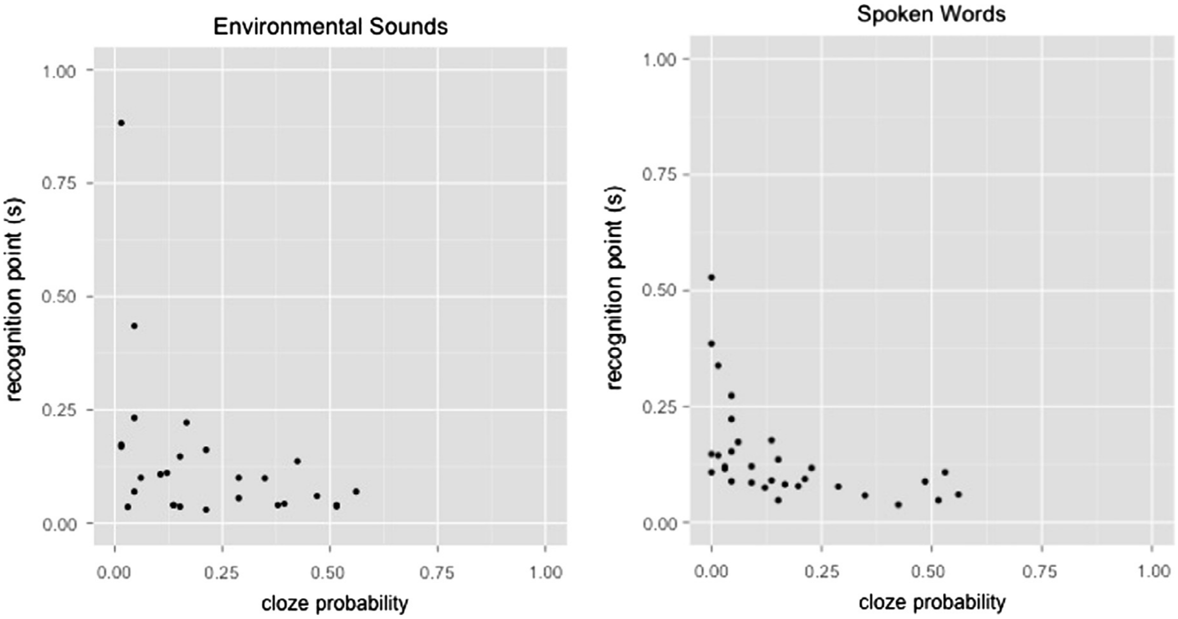

**Fig. 2.** There is an inverse nonlinear relationship between cloze probability of the general sentence frame and recognition point of the target following the sentence frame for both sounds and words.

typically, so there is an aspect of unnaturalness to the gating study. Furthermore, listeners in a gating study are not typically pressured to respond as quickly as possible, although Tyler and Wessels (1985) did not find that a speeded naming version of the gating task changed the results. Thus it has been claimed that gating is not qualitatively different from typical speech perception (Cotton & Grosjean, 1984; Tyler & Wessels, 1985, although see Allopenna, Magnuson, & Tanenhaus, 1998 for evidence that the gating paradigm may distort typical perception). Given that listeners presented with environmental sounds were not also identifying spoken word targets (between-subjects design), it seems unlikely that the environmental sound recognition was influenced by a specific word recognition strategy. Yet, it is also the case that listeners may use a more cognitive-inferential approach given the nature of the gating task than would be the case in normal speech perception. Furthermore, the gating task is an identification task rather than a meaning comprehension task. To address these concerns, we designed Experiment 2 using fluent speech and a comprehension task.

## 3. Experiment 2

To test understanding of the meaning of the sounds in sentence context, subjects were instructed to determine whether sentences were "understandable" or "nonsense". Half of the sentences were created to be understandable (the last word or environmental sound matched fit with the meaning of the sentence) and half were nonsense (the last word or its matched sound was highly implausible in the context of the sentence). Additionally, half of the stimuli ended in a spoken word target and half ended in an environmental sound target. If the similarity of sentence frame context effects for speech and non-speech in the first experiment were governed by a slower problem-solving strategy rather than a more fluent perceptual understanding process, speeded processing of spoken sentences should show a different pattern from the gating task.

### 3.1. Methods

#### 3.1.1. Participants

Participants were selected from the same population as in Experiment 1, but individuals were excluded from participating in both experiments to avoid effects from repeated exposure to the stimuli. There were 31 participants (15 female, mean: 20.93 years, range: 18–38 years). Informed consent was obtained and participants were paid at the same rate as in Experiment 1.

#### 3.1.2. Stimuli

The stimuli were taken from the same set as the previous study, but additional "nonsense" sentences were created as distractors. These were constructed by rearranging the ending words of the sentences. The resulting sentences were verified in a short written survey to ensure that they were not easily construed to make sense. For example, the sentence "He closed his winter jacket with the zipper," which is understandable, might be rearranged with the target "train" to form the nonsense sentence "He closed his winter jacket with the train." Thus, the "understandable" nature of the sentence depended on the last word of the sentence, which was replaced by a categorically matched environmental sound for half the stimuli. This 2 × 2 × 2 design gave rise to eight possible types of sentences: general/specific constraint × word/sound target × understandable/nonsense (all sentences are available in the Supplement).

Stimuli were presented at 65–70 dB over headphones. The experiment was coded in Matlab 2014 with Psychtoolbox 3.

#### 3.1.3. Task

The participants' task was to decide, as quickly and accurately as possible, whether the sentences were "understandable" or "nonsense". They responded by pressing one of two labeled keys, the side (right or left) of which was counterbalanced across participants. This task was chosen, as opposed to a sound recognition task, because recognition could require participants to "name" the sound. For this reason, we used an understandable/nonsense judgment rather than a sound recognition task (as in Potter et al., 1986) in order to judge if environmental sounds might also convey meaning without a naming step.

Each session of the experiment consisted of 5 blocks. The first block contained 20 practice spoken sentences, all ending in spoken words, for task familiarization. There was a 2 × 2 constraint × meaning design in the practice block, such that general-specific and understandable-nonsense conditions were equally represented. The remaining four experimental blocks consisted either of 32 spoken sentences with targets as environmental sounds (sound blocks) or of 32 sentences with targets as spoken words (word blocks). Half of the participants received blocks in a sound-word-sound-word order, and half the participants in word-sound-word-sound. Within each block, each target occurred exactly once. Targets never appeared in the same context more than once. For example, if "sheep" was heard in a general-frame, understandable context in block 1, it might appear in a general-frame, nonsense context in block 2, and a specific-frame, understandable context in block 3. Within each block, the order of stimuli was randomized. Participants received short breaks between blocks.

### *3.1.4. Data collection and analysis*

Response times (RTs) were recorded in Matlab with Psychtoolbox 3. RTs were defined as the time between the onset of the sentence's last word/sound ("target onset") and the participant's button press. Responses were classified as either (1) correct (the person responded with "understandable" or "nonsense" as appropriate after the onset of the last word or the sound), (2) incorrect (the person assigned the wrong understandability status to the sentence after the onset of the last word) or (3) guessing (the person responded before the onset of the last word, yielding a negative RT).

Three subjects were excluded, one due to low English proficiency and two for failure to follow instructions. Of these two, one was excluded for performance below 80% correct in the general/understandable/sounds condition, and one was excluded for excessive guessing, evidenced by negative RTs in 66% (general/nonsense/sounds) and 80% (specific/nonsense/sounds) of trials. Only understandable sentences with correct responses (i.e., understandable sentences to which participants correctly responded "understandable") were included in further analysis of RTs; nonsense sentences were treated as distractors and their RTs were not analyzed further. Incorrect trials and trials where participants guessed (negative RTs indicate responses before the target onset) were also excluded from further analysis.

A Repeated Measures ANOVA was performed on RT data. The factors modeled included constraint level of frame (general or specific), and target type (word or environmental sound), both within-subjects. As ANOVAs are not well-suited to modeling categorical data, we used a logit mixed model approach for the percent correct data (Jaeger, 2008). Using the *lme4* package in R (Bates, Maechler, Bolker, & Walker, 2015), we compared model A (main effects of constraint and target type, with random intercepts for subjects and sentence endings) with model B (model A plus constraint $*$ target type interaction). There was no evidence that including the interaction term improved the performance of the model ($\chi^2$ (1) = 0.028, $p > .8$); therefore we performed further analyses with model A.

### *3.2. Results*

Accuracy was high in all conditions, close to 90% or higher (Fig. 3).

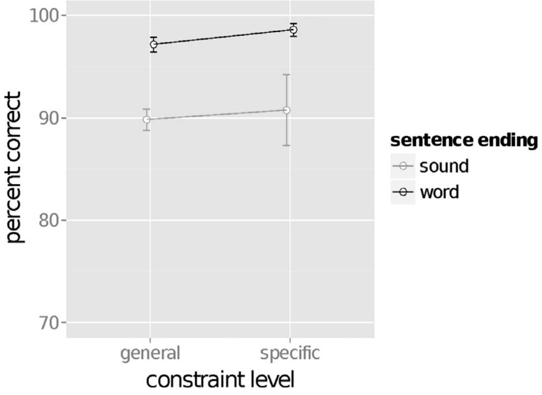

**Fig. 3.** Percent correct responses for meaningful vs nonsense judgments about sentences ending in either words or sounds. Accuracy was high (≥90%) for all conditions but higher for words than sounds. Mean percent correct ± SEM error bars shown.

**Table 2**
Coefficients for factors in a logit mixed model for percent correct data.

| Factor | Coeff | SE | Z | p |
|---|---|---|---|---|
| Intercept | 4.26 | 0.41 | 10.36 | 2E − 16 |
| Target type | − 1.45 | 0.47 | − 3.10 | 0.002 |
| Constraint | 0.66 | 0.25 | 2.62 | 0.009 |

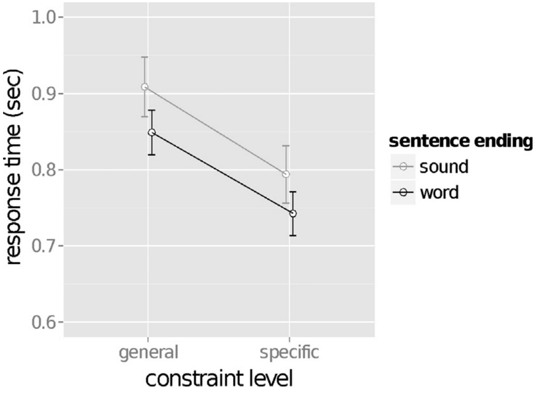

**Fig. 4.** RTs for meaningful vs nonsense judgment. Response time measured from target (i.e. last item in sentence) onset shows that sentence constraint speeds meaningfulness responses similarly for stimuli ending in spoken words and environmental sounds. Mean RT ± SEM error bars shown.

There was significantly higher accuracy in specific (compared to general) context conditions (p = .009, Table 2); and significantly higher accuracy in word (compared to environmental sound) conditions (p = .002, Table 2). There was no evidence for an interaction between these terms.

For response times, there was a main effect of sentence constraint, such that meaningfulness judgments for specific (i.e. high constraint) sentences were faster than general (i.e. low constraint) sentences (Fig. 4, mean RTs 768 vs 879 ms, $F$ (1, 27) = 52.34, $p < .001$, $d$ = 2.79). The main effect of target type (i.e. spoken word versus nonspeech sound) approached significance ($F$ (1, 27) = 3.67, $p$ = .066, $d$ = 0.74), but the interaction of target type with constraint level was not significant ($p > .25$).

## 4. General discussion

Listeners have little difficulty understanding spoken sentences that end in "sound effects" that substitute for spoken words. Though accuracy for all-word sentences was higher than for sentences ending in nonspeech environmental sounds, overall accuracy was quite high in both conditions at 90% correct or above, suggesting that even such an unusual, unfamiliar task as understanding environmental sounds in sentence context is not much more difficult than the everyday task of understanding a normal sentence. How can we explain this effect? It is unlikely that listeners use a covert naming strategy—naming in other studies takes several hundred milliseconds, and we find no delays of this magnitude. In Experiment 2, RTs were on average approximately 60 ms slower for environmental sounds, and this difference did not reach significance. It is difficult to justify any model of language understanding that depends on a dedicated speech processor given that nonspeech can be understood in spoken sentences and derive the same contextual benefit as spoken words. In light of the current results, a purely lexical model (e.g., the distributed Cohort model Gaskell & Marslen-Wilson, 1997) cannot explain the results without substantial and fundamental changes to the underlying assumptions. Of course, environmental sounds, like spoken words, could be treated as meaning-associated patterns in the same system albeit with a different kind of feature base and pattern processing system Tyler, Voice, and Moss

(1996) suggested that the top tier of the nodes in the TRACE model (i.e., words) could be linked to an even higher layer, containing nodes that represent concepts or semantic information; context effects could occur by activity in this layer feeding back onto the lower levels. If there is a separate network of neural representations of environmental sounds that is also one layer below the conceptual layer, then activity from sentence context in the conceptual layer could feed back onto the sound recognition layer, affecting recognition.

The present experiments demonstrate that sentential context can provide equal benefit for both spoken words and non-speech sounds in recognition of the targets as well as in understanding a whole sentence, as if the non-speech sound is a natural part of the sentence. From a predictive coding perspective, this suggests that even in situations where it is not possible to make neural predictions via motor systems (as our environmental sounds do not have clear speech or other motor representations), some other type of predictions can constrain processing (cf Hickok, 2012). In other words, these results suggest that neural predictions can occur at a conceptual level in constraining recognition and comprehension. The acoustics of environmental sounds are more variable and less predictable than the acoustics of words from the same speaker that the participants heard in the sentence frame. Therefore, it follows that accurate sensory predictions for environmental sounds must be more difficult to form. Moreover, if predictive coding is highly statistically dependent as suggested by Kuperberg and Jaeger (2016), the statistical rarity of environmental sounds serving to complete sentence frames would pose a substantial challenge. If participants need to rely heavily on precisely tuned sensory predictions from context to constrain processing, we would not have found such similar context effects for both environmental sounds and words. Of course, this is not to say that sensory predictions do not contribute to language understanding; merely that conceptual predictions can be sufficient to constrain processing to much the same extent. Thus, a processing framework like that proposed by Lupyan and Clark (2015) could apply, with the caveat that the importance of the "low-level predictions" is variable based on the sensory quality or statistical properties of the stimuli. Future work could compare neural responses to environmental sounds and spoken words in sentence context Lewis and Bastiaansen (2015) theorize that low and high gamma-range oscillations represent propagation of top-down predictions and computation of prediction errors, respectively. The relative importance of these processes for recognizing and understanding environmental sounds versus spoken words in sentence context can be tested with our stimuli set and EEG time-frequency analyses.

While in many respects environmental sounds and words behaved similarly in our experiments, the 60 ms processing cost for environmental sounds was an important difference. It is unlikely that this difference is an artifact of the properties of the stimuli sets used, because there was no significant difference in amount of waveform needed for recognition of meaningful sounds and spoken words in Experiment 1. If the small RT difference observed for meaningfulness decisions for normal and rebus sentences was due to longer average recognition time for the environmental sounds than the words, we might expect systematic differences in recognition points to show up in the gating study, but none were apparent (i.e. no main effect of target type or context-target type interaction).

One explanation for this processing cost is that in order to switch from interpreting the experiment's male speaker to an environmental sound, listeners may have to shift attention in some sense. For example, when an object appears in an unexpected location or changes form unexpectedly, observers engage in shifting attention (Yantis & Serences, 2003). This is similar to the ~40 ms processing cost that is incurred when talkers change (Nusbaum & Morin, 1992; Nusbaum & Magnuson, 1997). Interestingly, music work suggests that a processing cost of about this size is not unique to language, but could reflect attention reallocation in other types of auditory processing as well. Van Hedger et al. (2015) found a processing cost near 40 ms for both switching between timbres and between octaves in a paradigm where absolute pitch possessors were asked to respond to certain target notes. Perhaps the additional 20 ms delay observed in our paradigm reflects switching attention to a farther-away modality, as both the talker change and timbre/octave change experiments took place in the same modality. Regardless, the need for attention reallocation did not impair participants' ability to make use of context information to speed processing. This suggests that signal source changes (e.g., timbre, voice, speech/nonspeech) can slow recognition but still engage the same processes that are also needed for sentence understanding (Heald & Nusbaum, 2014a).

Another potential source for the 60 ms environmental sound processing cost is the relative unfamiliarity of the sounds, particularly in sentence context. It may be the case that, on top of reallocation of attention, a processing lag is imposed by virtue of the environmental sounds being less frequently heard than spoken words, and almost never heard as meaningful items in a spoken sentence as they were in our experiment. This interpretation is supported by the large body of work on frequency effects in spoken word recognition (e.g. Dahan et al., 2001; Luce & Pisoni, 1998), as well as by Van Hedger et al. (2015), who found that responses were significantly slower to octaves with which the participants had less musical experience based on the instrument that they played. An effect of reduced familiarity can also be explained as a predictive coding disadvantage, as more commonly encountered stimuli are more likely to have stronger neural representations that can be activated as predictions. Perhaps because in this paradigm, it is easier to form sensory-level predictions for spoken words, the words can benefit from stronger low-level predictions that speed processing. It is interesting to note that this cost is only 60 ms, which implies that strong conceptual-level predictions are able to accommodate substantial variance in the input to the system—even when that input is no longer linguistic. In any case, it follows that if at least part of the sound-word difference is due to differential experience, training should be able to narrow this gap. Future experiments might train participants on a subset of environmental sounds by pairing them with pictures, and then compare participants' speed of integrating trained sounds versus words with sentence context in order to address this question. ERP analyses could also reveal differences in the time courses of sound and word processing that are too fine-grained to be picked up by behavioral studies such as the ones reported here. Such analyses could reveal whether differences between sounds and words are restricted to early, attention-capture stages (e.g. auditory evoked potentials like the N1/P2), whether they persist late in processing (e.g., N400, P600), and to what extent they might be able to explain a ~60 ms processing cost.

It is worth noting that many extant models of speech recognition do not explicitly account for how such seamless interactions between verbal and nonverbal stimuli might be happening in terms of a mechanism. From a perspective concerning how these models are used to explain language understanding, our results are interesting because they suggest some updates to these models. A number of different studies (Shintel & Nusbaum, 2007; Zwaan & Pecher, 2012; Zwaan et al., 2002) have demonstrated that language, specifically an intact clause referring to an object, can facilitate understanding of that object. In these cases, an understandable linguistic form (a word, a clause, or a sentence) refers to a non-linguistic object and speeds processing for that object. The present experiments go one step further, because a non-linguistic object is not referred to by a complete sentence, but is directly incorporated into the sentence as if it were itself a linguistic form. In other words, the sentences in our Experiment 2 do not function as complete ideas without incorporating the meanings of the environmental sounds. From the position of a general cognitive processing system in which linguistic forms hold no special or privileged status either by virtue of their high degree of statistical association or by virtue of specialized mechanisms subserving their processing, the present results are predicted and expected. Even without a claim of modularity however, connectionist models—e.g., Shortlist B, updates of

TRACE, interactive Hebbian models—that are used to model speech recognition would not fluently treat a non-speech sound as substitutable for recognition or understanding purposes (Mcclelland et al., 2006; Mirman et al., 2006; Norris & McQueen, 2008; Strauss et al., 2007). Learning models of speech (Jurafsky & Martin, 2000; Kuperberg & Jaeger, 2016; McMurray, Aslin, & Toscano, 2009) operate on associative statistical principles, and non-speech sounds do not occur in these contexts. While a general cognitive processing perspective can easily account for the rapid shift of attention to an unlikely sound if that sound's interpretation fits with the contextual meaning of the antecedent frame, most language models do not take this into account.

The present results demonstrate that there is no evidence for a differential effect of context for meaningful nonspeech sounds relative to matched spoken words. Whether the meaning is derived from a vocal-tract produced utterance, or from environmental generators, context appears to similarly limit recognition and understanding. These results highlight the importance of conceptual meaning in context effects, as these two types of signals are vastly different in their acoustic properties, sources, and statistical occurrences, but are well matched in terms of conceptual meaning.

In summary, our results have strong implications for language processing theories. They add to the considerable body of evidence arguing against modular, encapsulated language processing by demonstrating that understanding words and environmental sounds in the same sentence requires an attention switch akin to switching between two talkers, rather than a deductive or covert naming strategy. Moreover, our results suggest that, if predictive coding is responsible for the facilitative effects of constraint on language processing, it is likely that predictions involving general conceptual representations (as opposed to low-level sensory predictions), are largely sufficient to drive constraint effects. Finally, our results suggest that models of speech comprehension that largely rely on lexical attributes should be modified to include a larger contribution from general cognitive processes that take conceptual meaning into account.

## Acknowledgements

The authors would like to thank Bruno Cabral, Alison Trude, Jean Matelski-Boulware, Heather Mangelsdorf, Nina Bartram, and Peter Hu for help with creation of the stimuli and piloting of the experiments. They also thank Brendan Colson and Michael Stubna for data collection, and Willow Uddin-Riccio for help preparing the manuscript.

This research was supported by the Multidisciplinary University Research Initiatives (MURI) Program of the Office of Naval Research through grant, DOD/ONR N00014-13-1-0205, and by the University of Chicago MSTP Training Grant (T32GM007281).

## Appendix A. List of Stimuli

See Table A1.

**Table A1**
Paired environmental sounds and spoken words used in the current study.

| Sound | Word |
|---|---|
| Baby laughing | "baby" |
| Camera shutter | "camera" |
| Car engine revving | "car" |
| Cashregister ch-ching | "cashregister" |
| Cat meowing | "cat" |
| Churchbells ringing | "churchbells" |
| Clock ticking | "clock" |
| Coin dropping onto hard surface | "coin" |
| Cow mooing | "cow" |
| Crow cawing | "crow" |
| Dog barking | "dog" |
| Creaky door closing | "door" |
| Doorbell ringing | "doorbell" |
| Drum set | "drums" |
| Frog croaking | "frog" |
| Guitar being strummed | "guitar" |
| Gunshot | "gunshot" |
| Helicopter | "helicopter" |
| Car horn | "horn" |
| Papers being ruffled | "paper" |
| Phone ringing | "phone" |
| Octave played on piano | "piano" |
| Rooster crowing | "rooster" |
| Saxophone notes | "saxophone" |
| Servicebell ringing | "servicebell" |
| Sheep bleating | "sheep" |
| Ambulance/police siren | "siren" |
| Sword being unsheathed | "sword" |
| Toilet flushing | "toilet" |
| Train whistle | "train" |
| Water dripping | "water" |
| Zipper | "zipper" |

## Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2017.12.009.

## References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439. http://dx.doi.org/10.1006/jmla.1997.2558.

Ballas, J. A., & Mullins, T. (1991). Effects of context on the identification of everyday sounds. *Human Performance, 4*, 199–219. http://dx.doi.org/10.1207/s15327043hup0403_3.

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

Bonhage, C. E., Mueller, J. L., Friederici, A. D., & Fiebach, C. J. (2015). Combined eye tracking and fMRI reveals neural basis of linguistic predictions during sentence comprehension. *Cortex, Special Issue: Prediction in Speech and Language Processing, 68*, 33–47. http://dx.doi.org/10.1016/j.cortex.2015.04.011.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

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. http://dx.doi.org/10.1037/0278-7393.30.3.687.

Chen, Y.-C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 1554–1568. http://dx.doi.org/10.1037/a0024329.

Cotton, S., & Grosjean, F. (1984). The gating paradigm: A comparison of successive and individual presentation formats. *Perception & Psychophysics, 35*, 41–48.

Cummings, A., Čeponienė, R., Koyama, A., Saygin, A. P., Townsend, J., & Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain Research, 1115*, 92–107. http://dx.doi.org/10.1016/j.brainres.2006.07.050.

Dahan, D., & Magnuson, J. S. (2006). Chapter 8 – Spoken Word Recognition A2 - Traxler, Matthew J. In M. A. Gernsbacher (Ed.). *Handbook of psycholinguistics* (pp. 249–283). (second ed.). London: Academic Press.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42*, 317–367. http://dx.doi.org/10.1006/cogp.2001.0750.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 30*, 498–513. http://dx.doi.org/10.1037/0278-7393.30.2.498.

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. http://dx.doi.org/10.1038/nn1504.

Dick, F., Krishnan, S., Leech, R., Saygin, A. P. (2016). Environmental sounds. In *Neurobiology of Language* (pp. 1121–1138). Elsevier.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*, 469–495. http://dx.doi.org/10.1006/jmla.1999.2660.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75–84. http://dx.doi.org/10.1016/j.brainres.2006.06.101.

Frey, A., Aramaki, M., & Besson, M. (2014). Conceptual priming for realistic auditory scenes and for auditory words. *Brain and Cognition, 84*, 141–152. http://dx.doi.org/10.1016/j.bandc.2013.11.013.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*, 613–656.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception, & Psychophysics, 28*, 267–283. http://dx.doi.org/10.3758/BF03204386.

Heald, S. L. M., & Nusbaum, H. C. (2014a). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*. http://dx.doi.org/10.3389/fnsys.2014.00035.

Heald, S. L. M., & Nusbaum, H. C. (2014b). Talker variability in audio and audiovisual speech perception. *Frontiers in Psychology, 5*, 698. http://dx.doi.org/10.3389/fpsyg.2014.00698.

Hedger, S. C., Nusbaum, H. C., & Hoeckner, B. (2013). Conveying movement in music and prosody. *PLoS ONE*. http://dx.doi.org/10.1371/journal.pone.0076744.

Hickok, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. J. Commun. Disord., 21st Annual NIDCD-Sponsored ASHA Research Symposium (2011): Neuroplasticity in the Mature Brain 45, 393–402. http://dx.doi.org/10.1016/j.jcomdis.2012.06.004.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., ... Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods, 45*, 1099–1114. http://dx.doi.org/10.3758/s13428-012-0304-z.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language, 59*, 434–446. http://dx.doi.org/10.1016/j.jml.2007.11.007.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (second ed.). London: Pearson.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception, 36*, 1–16.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*, 32–59. http://dx.doi.org/10.1080/23273798.2015.1102299.

Leech, R., & Saygin, A. P. (2011). Distributed processing and cortical specialization for speech and environmental sounds in human temporal cortex. *Brain and Language, 116*, 83–90. http://dx.doi.org/10.1016/j.bandl.2010.11.001.

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex, 68*, 155–168. http://dx.doi.org/10.1016/j.cortex.2015.02.014.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*, 1–36.

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science, 24*, 279–284. http://dx.doi.org/10.1177/0963721415570732.

Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition, 108*, 866–873. http://dx.doi.org/10.1016/j.cognition.2008.06.005.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86. http://dx.doi.org/10.1016/0010-0285(86)90015-0.

Mcclelland, J., Mirman, D., & Holt, L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10*, 363–369. http://dx.doi.org/10.1016/j.tics.2006.06.007.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science, 12*, 369–378. http://dx.doi.org/10.1111/j.1467-7687.2009.00822.x.

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition, 33*, 1174–1184. http://dx.doi.org/10.3758/BF03193221.

Meteyard, L., Bahrami, B., & Vigliocco, G. (2007). Motion detection and motion verbs: Language affects low-level visual perception. *Psychological Science, 18*, 1007–1013. http://dx.doi.org/10.1111/j.1467-9280.2007.02016.x.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language, 66*, 545–567. http://dx.doi.org/10.1016/j.jml.2012.01.001.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227–234. http://dx.doi.org/10.1037/h0031564.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13*, 958–965.

Morris, A. L., & Harris, C. L. (2002). Sentence context, word recognition, and repetition blindness. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 28*, 962–982. http://dx.doi.org/10.1037//0278-7393.28.5.962.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*, 189–234. http://dx.doi.org/10.1016/0010-0277(94)90043-4.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*, 357–395. http://dx.doi.org/10.1037/0033-295X.115.2.357.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. *Talker Variability in Speech Processing*, 109–132.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.). *Speech perception, production, and linguistic structure* (pp. 113–134). Tokyo: OHM Publishing Company.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology, 17*, 273–281. http://dx.doi.org/10.1080/17470216508416445.

Orgs, G., Lange, K., Dombrowski, J.-H., & Heil, M. (2006). Conceptual priming for environmental sounds and words: An ERP study. *Brain and Cognition, 62*, 267–272. http://dx.doi.org/10.1016/j.bandc.2006.05.003.

Orgs, G., Lange, K., Dombrowski, J., & Heil, M. (2007). Is conceptual priming for environmental sounds obligatory? *International Journal of Psychophysiology, 65*, 162–166. http://dx.doi.org/10.1016/j.ijpsycho.2007.03.003.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*, 105–110. http://dx.doi.org/10.1016/j.tics.2006.12.002.

Pickering, M. J., & Garrod, S. (2013). Forward models and their implications for production, comprehension, and dialogue. *Behavioral and Brain Sciences, 36*(4), 377–392.

Potter, M. C., Kroll, J. F., Yachzel, B., Carpenter, E., & Sherman, J. (1986). Pictures in sentences: Understanding without words. *Journal of Experimental Psychology: General, 115*, 281.

Revill, K. P., Aslin, R. N., Tanenhaus, M. K., & Bavelier, D. (2008). Neural correlates of partial lexical activation. *Proceedings of the National Academy of Sciences, 105*, 13111–13115. http://dx.doi.org/10.1073/pnas.0807054105.

Saygin, A. P. (2003). Neural resources for processing language and environmental sounds: Evidence from aphasia. *Brain, 126*, 928–945. http://dx.doi.org/10.1093/brain/awg082.

Saygin, A. P., Dick, F., & Bates, E. (2005). An on-line task for contrasting auditory processing in the verbal and nonverbal domains and norms for younger and older adults. *Behavior Research Methods, 37*, 99–110.

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental Psychology, 55*, 121–132. http://dx.doi.org/10.1027/1618-3169.55.2.121.

Shillcock, R. C., & Bard, E. G. (1993). Modularity and the processing of closed-class words. *Cognitive models of speech processing: The second sperlonga meeting* (pp. 163–185). East Sussex: Lawrence Erlbaum Associates Ltd.

Shintel, H., & Nusbaum, H. C. (2007). The sound of motion in spoken language: Visual information conveyed by acoustic properties of speech. *Cognition, 105*, 681–690. http://dx.doi.org/10.1016/j.cognition.2006.11.005.

Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language, 55*, 167–177. http://dx.doi.org/10.1016/j.jml.2006.03.002.

Simpson, G. B., Peterson, R. R., Casteel, M. A., & Burgess, C. (1989). Lexical and sentence context effects in word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 15*, 88–97. http://dx.doi.org/10.1037/0278-7393.15.1.88.

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82*, 1–17. http://dx.doi.org/10.1016/j.jml.2015.02.004.

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). JTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods, 39*, 19–30.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645–659. http://dx.doi.org/10.1016/S0022-5371(79)90355-4.

Tanenhaus, MK., Spivey-Knowlton, MJ., Eberhard, KM., & Sedivy, JC. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*.

Tyler, L. K., Voice, J. K., Moss, H. E. (1996). The interaction of semantic and phonological processing. In *Eighteenth annual conference of the cognitive science society*.

Tyler, L. K., & Marslen-Wilson, W. (1986). The effects of context on the recognition of polymorphemic words. *Journal of Memory and Language, 25*, 741–752. http://dx.doi.org/10.1016/0749-596X(86)90047-1.

Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics, 34*, 409–420.

Tyler, L. K., & Wessels, J. (1985). Is gating an on-line task? Evidence from naming latency data. *Perception, & Psychophysics, 38*, 217–222. http://dx.doi.org/10.3758/BF03207148.

Van Hedger, S. C., Heald, S. L. M., & Nusbaum, H. C. (2015). The effects of acoustic variability on absolute pitch categorization: Evidence of contextual tuning. *Journal of the Acoustical Society of America, 138*, 436–446. http://dx.doi.org/10.1121/1.4922952.

van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia, 33*, 485–508. http://dx.doi.org/10.1016/0028-3932(94)00133-A.

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16*, 1173–1184.

Yantis, S., & Serences, J. T. (2003). Cortical mechanisms of space-based and object-based attentional control. *Current Opinion in Neurobiology, 13*(2), 187–193. http://dx.doi.org/10.1016/S0959-4388(03)00033-3.

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PLoS ONE, 7*, e51382. http://dx.doi.org/10.1371/journal.pone.0051382.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science, 13*, 168–171. http://dx.doi.org/10.1111/1467-9280.00430.