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