Journal of Memory and Language 100 (2018) 32–50

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Is speech recognition automatic? Lexical competition, but not initial lexical access, requires cognitive resources

Xujin Zhang[a], Arthur G. Samuel[a,b,c,*]

[a] Psychology Department at Stony Brook University, 100 Nicolls Road, Stony Brook University, Stony Brook, NY, 11794-2500, USA
[b] Basque Center on Cognition Brain and Language, Paseo Mikeletegi 69, 2nd Floor, 20009, Donostia-San Sebastián, Gipuzkoa, Spain
[c] IKERBASQUE Basque Foundation for Science, Alameda Urquijo, 36-5, Plaza Bizkaia, 48011 Bilbao, Bizkaia, Spain

ARTICLE INFO

Keywords:
Speech recognition
Lexical access
Lexical competition
Cognitive load
Cognitive resources

ABSTRACT

Current models of spoken word recognition suggest that multiple lexical candidates are activated in parallel upon hearing an utterance, with these lexical hypotheses competing with each other for recognition. The current project investigated the effect of cognitive load on initial lexical access and later lexical competition. In a set of priming studies, the lexicality of the primes (i.e., non-word vs. word) was manipulated to dissociate these two sub-processes. We tested performance on a semantic association task under conditions with no additional load, or with cognitive load that used cognitive resources that are either general or more specific to phonological processing. The results suggest that the initial access of lexical items is relatively automatic. In contrast, maintaining lexical candidates in competition requires cognitive resources, and these resources are specific to phonological processing. The overall result pattern provides insights into differences in the way that lexical activation and competition operate.

Introduction

Understanding spoken language is one of the most fundamental cognitive skills human beings have. Speakers first formulate semantic information they would like to express, select proper lexical items, activate the phonological information for these items, and use the motor system to articulate sounds. Listeners map the acoustic-phonetic waveform of the unfolding speech signal to the lexical representations stored in long-term memory, find the right item in long-term memory, activate its semantic representation, and understand a spoken word. For normal adults, speech recognition is fast and seems effortless, but the underlying mechanism is complex. A critical question is whether speech recognition is as automatic as we subjectively feel. In the current study, we compare speech recognition under optimal vs. more difficult conditions to test which sub-processes during speech recognition really do operate relatively automatically, and which require cognitive resources.

Lexical access and competition

Decades of research have been devoted to the question of how spoken words are recognized with such remarkable efficiency. Most current models of spoken word recognition (Cohort: Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978; TRACE: McClelland & Elman, 1986; Shortlist: Norris, 1994) agree that when speech comes in, the signal first makes contact with sub-lexical representations, such as acoustic-phonetic features or phonemes. The processing at the sub-lexical level provides input codes for accessing lexical entries, where the form (e.g., abstract phonological information, morphological information), syntactic role, and semantic information of words are stored. Although different models make different claims about the dynamic properties of speech processing, there is a consensus that upon hearing the first few segments of an unfolding speech signal, multiple lexical entries are activated automatically in parallel if their phonological representations transiently match the incoming signal. This initial lexical access is thought to occur as early as the first 100–150 ms of a speech signal, and to occur obligatorily. The bottom-up activation of a lexical candidate depends merely on the goodness-of-fit between the speech signal and the phonological representation of the candidate.

There have been a large number of studies supporting rapid initial access of multiple lexical candidates. Various tasks have been used, including gating (e.g., Grosjean, 1980), shadowing (e.g., Marslen-Wilson, 1973), perceptual identification (e.g., Slowiaczek, Nusbaum, & Pisoni, 1987), lexical decision (e.g., Goldinger, Luce, Pisoni, & Marcario, 1992; Zwitserlood,1989), word spotting (e.g., McQueen, Norris, & Cutler, 1994), eye-tracking (e.g., Allopenna, Magnuson, & Tanenhaus, 1998), and ERPs (e.g., Friedrich, Felder, Lahiri, & Eulitz,

* Corresponding author at: Department of Psychology, Stony Brook University, NY 11794-2500, USA.
E-mail addresses: xujin.zhang@stonybrook.edu (X. Zhang), arthur.samuel@stonybrook.edu, a.samuel@bcbl.eu (A.G. Samuel).

https://doi.org/10.1016/j.jml.2018.01.002
Received 31 December 2015; Received in revised form 3 January 2018

2013). There is also substantial evidence that multiple lexical access occurs obligatorily, regardless of contextual constraints. For instance, even when the semantic or syntactic context favors only one of the lexical hypotheses, all possible candidates are activated before the uniqueness point of a spoken word is heard (e.g., Dahan & Tanenhaus, 2004; Zwitserlood, 1989). Similarly, all possible meanings of a polysemous word and all possible interpretations of a homophone or ambiguous-sounding word are activated at the beginning of the speech, independent of the context (e.g., Connine, Blasko, & Wang, 1994; Swinney, 1979; Tanenhaus, Leiman, & Seidenberg, 1979).

The literature on visual word recognition provides additional evidence that mapping sensory information onto lexical representations occurs automatically, without intention and awareness (see Neely, 1991 for a review). For instance, words are activated to the level of meaning even when participants are instructed to ignore them (e.g., Fuentes, Carmona, Agis, & Catena, 1994), when participants' attention is allocated to lower-level information rather than the meaning of the words (e.g., letters: Valdés, Catena, & Marí-Beffa, 2005; ink color: MacLeod, 1991; MacLeod, 1992; Stroop, 1935), and even when participants are not consciously aware of the presence of the words (e.g., Marcel, 1983). These results indicate that initial lexical access based on bottom-up activation functions in a relatively automatic way and may not require much attentional control.

According to current models of spoken word recognition (Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Norris, 1994), once multiple lexical hypotheses are generated by the speech signal, a competition mechanism is necessary for the selection of the best candidate to be recognized. One type of competition depends on the degree of match or mismatch between the bottom-up signal and the phonological representations of lexical candidates. The Cohort model (Marslen-Wilson, 1987; Marslen-Wilson, Moss, & Van Halen, 1996) assumes that the activation level of a candidate is reduced when the unfolding speech input is no longer consistent with it. For instance, although for Dutch listeners, both "kapitein" and "kapitaal" are activated upon hearing "kapit", once the vowel after "t" is heard, responses to a probe associated with the other candidate are no longer facilitated (Zwitserlood,1989). However, this does not mean that the mismatching candidate is completely eliminated from the candidate set or is excluded from future processing. Dahan and Gaskell (2007) found that although fixations to a cohort competitor decreased after the recognition point of the target word, they were still greater than those to unrelated distracters. In addition, studies of embedded words have also shown robust priming for the embedded words (e.g., "cap" within "captain") at the offset of (Isel & Bacri, 1999; Luce & Cluff, 1998; Vroomen & De Gelder, 1997), 100 ms after (Macizo, Van Petten, & O'Rourke, 2012), and 500 ms after the carrier words (Zhang & Samuel, 2015), suggesting an extended time window of activation (Dahan & Gaskell, 2007; Friedrich et al., 2013; Marslen-Wilson & Welsh, 1978).

Another type of competition comes from co-activated lexical candidates. The TRACE model (McClelland & Elman, 1986) assumes that activated candidates compete directly with each other via lateral inhibition. All activated candidates inhibit each other as a function of their bottom-up activation level, which depends on their similarity to the speech signal. At any time during perception, the activation level of a candidate is determined by the bottom-up activation received from the speech input and the lateral inhibition received from other activated candidates. The candidate that is most similar to the speech signal usually has the strongest activation and sends out the strongest inhibition to other candidates, and therefore will win the competition. Furthermore, short words usually have a disadvantage over long words because short words receive less bottom-up support from the speech input than longer words, and they receive more competition from similar sounding words (Bowers, Davis, Mattys, Damian, & Hanley, 2009).

No studies have explicitly examined whether lexical competition is as automatic as initial lexical access. However, some studies have suggested that distinguishing among lexical candidates and inhibiting inappropriate ones may take more time than activating those candidates (e.g., Marslen-Wilson, 1987; Swinney, 1979) and may be relatively costly in terms of processing resources (e.g., Connine et al., 1994). Moreover, research on language deficits has also indicated that processes such as inhibition might be more likely to vary between individuals than activation (e.g., McMurray, Samelson, Lee, & Tomblin, 2010). Across the different views of lexical competition, a common feature is the need to maintain the competing candidates themselves during the competition, which itself may be resource-dependent.

Collectively, the available evidence suggests that initial lexical access and later lexical competition – two sub-processes involved in speech recognition – may have different requirements for cognitive resources and attentional control. However, as noted, there has not been explicit investigation of the automaticity of lexical access versus competition. The current study addresses this issue by comparing initial lexical access and later lexical competition under both optimal and more complicated conditions. In the latter, cognitive resources were depleted by secondary cognitive load tasks.

Effect of cognitive load on speech processing

There has been a recent growth in work focusing on speech perception under more complicated situations. For instance, studies have examined speech perception under perceptual load due to background noise or changed speaking rates, or under cognitive load imposed by secondary tasks (see Mattys, Davis, Bradlow, & Scott, 2012 for a review). Cognitive load research has shown that speech is sometimes processed in the same way under optimal conditions as under cognitive load, while sometimes not, implying that some processes during speech perception depend on the availability of cognitive resources more than others. For instance, the speech system is able to adjust to atypical pronunciations (Eisner & McQueen, 2005; Eisner & McQueen, 2006; Kraljic & Samuel, 2005; Kraljic & Samuel, 2006; McQueen, Cutler, & Norris, 2006; McQueen et al., 2006; Norris, McQueen, & Cutler, 2003) and to perceptually restore missing phonemes (Samuel, 1981; Samuel, 1996; Warren, 1970) under optimal conditions, and these abilities remain almost intact under cognitive load conditions (Mattys, Barden, & Samuel, 2014; Zhang & Samuel, 2014). However, for speech segmentation, listeners' reliance on fine-grained acoustic detail is attenuated under cognitive load (Mattys, Brooks, & Cooke, 2009; Mattys, Carroll, Li, & Chan, 2010).

In addition, previous studies have found that under optimal conditions, carrier words are able to prime words that are associated with words embedded in them (Bowers et al., 2009; Salverda, Dahan, & McQueen, 2003; Van Alphen & Van Berkum, 2010; Zhang & Samuel, 2015). However, when a cognitive load task is added, the carrier words (e.g., "napkin") no longer prime the associations (e.g., "sleep") of embedded words (e.g., "nap"), whereas the isolated embedded words (i.e., "nap") are still able to produce significant associative priming (Zhang & Samuel, 2015). These results indicate that cognitive load does not prevent the speech input from activating the meaning of a candidate, if its phonological representation perfectly matches the speech. The null effect for embedded words when hearing carrier words under cognitive load suggests that the consideration of lexical candidates that do not strongly match the speech is resource-dependent.

There are two possible explanations for this pattern. One is that cognitive load prevents alternative candidates from being accessed in the first place, which would occur if the initial lexical access based on bottom-up activation requires cognitive resources. Under optimal conditions, when there is no cognitive load, all possible candidates that match the speech signal to some degree can be activated at the same time. Although there is competition from the inconsistent bottom-up signal and/or from other candidates, the residual activation of some alternative candidates is still strong enough to be observed at the end of the speech input. In contrast, when processing demand increases, e.g.,