The Time Course of Interpretation in Speech Comprehension

Current Directions in Psychological Science
19 (2) 121-126

Delphine Dahan
University of Pennsylvania

Abstract
Determining how language comprehension proceeds over time has been central to theories of human language use. Early research on the comprehension of speech in real time put special emphasis on the sequential property of speech, by assuming that the interpretation of what is said proceeds at the same rate that information in the speech signal reaches the senses. The picture that is emerging from recent work suggests a more complex process, one in which information from speech has an immediate influence while enabling later-arriving information to modulate initial hypotheses. "Right-context" effects, in which the later portion of a spoken stimulus can affect the interpretation of an earlier portion, are pervasive and can span several syllables or words. Thus, the interpretation of a segment of speech appears to result from the accumulation of information and integration of linguistic constraints over a larger temporal window than the duration of the speech segment itself. This helps explain how human listeners can understand language so efficiently, despite massive perceptual uncertainty in the speech signal.

Keywords
speech comprehension, spoken-word recognition, sequential analysis, eye movements.

Watching closed captions—-the transcription of the audio portion of a program—on a television in a noisy airport provides for an effective demonstration of the complexity associated with comprehending speech in real time. Speech is a highly complex, fleeting stimulus, with two to three words produced per second. Captioning systems often cannot keep up with this pace, and the visual portion of the speech and its written transcription quickly fall out of sync. Moreover, the transcription of the initial portion of an utterance is often tentative and needs to be modified in the light of what comes later, with time-consuming revisions. This state of affairs contrasts with our experience of listening to the audio part of the same program, where understanding what people mean as they talk seems fluid and effortless.

Determining how we achieve this feat received little attention until the groundbreaking work of Cole, Marslen-Wilson, and their collaborators (e.g., Cole, 1981; Marslen-Wilson & Tyler, 1980), whose studies suggested that the perceptual interpretation of continuous speech occurs in real time, with linguistic analysis accomplished at the rate at which information arrives to the senses. In more recent years, however, a substantially different view of speech perception has emerged, one in which processing is seen as both rapid and flexible and in which information from speech is seen as having an immediate influence while also enabling later-arriving information to modulate initial hypotheses.

Corresponding Author:
Delphine Dahan, University of Pennsylvania, Dept. of Psychology, Rm302C
3401 Walnut Street, Philadelphia, PA 19104
Email: dahan@psych.upenn.edu

Speech Is Interpreted in Real Time

How can listeners keep up with the speed at which information arrives? Marslen-Wilson (1975) showed that people who were asked to "shadow" continuous speech, rapidly repeating back sentences as they heard them, often spontaneously corrected mispronunciations that had been placed at the ends of words. For example, people restored the mispronounced word "tomorrane" to its original form "tomorrow." The speed of these restorations implied that people had anticipated which word they were hearing based on its first sounds and planned their own production of the word even before the mispronunciation was noticed. Thus, hearing the early fragment of a word can elicit the percept of the word because people bring their knowledge of what could be said to the analysis of what is actually said from the earliest moments.

This and related findings led to the development of theories that assume that speech is submitted to a sequential, "left to right" analysis: Early information in the speech signal is utilized extremely rapidly and effectively to constrain the set of possible interpretations, in effect to predict what the upcoming signal may be—arguably a key aspect of real-time efficiency.

Much of the initial evidence for the sequential analysis of speech came from analyses of people's responses when asked to decide which word they were hearing upon listening to just its beginning parts (Grosjean, 1980). Often, listeners converged on the intended word without hearing it in its entirety. More recently, examination of listeners' eye movements to real or pictured objects as they hear instructions to manipulate one of them has confirmed the immediate uptake of information from the speech signal (see Fig. 1). Analyses of gaze location have revealed that, upon hearing the name of a referent object (e.g., "tap"), people are more likely to briefly fixate on a distractor object with a name that begins with the same sounds as that of the referent (a competitor, e.g., "tack") than on a distractor with an unrelated name (e.g., "deer"; Allopenna, Magnuson, & Tanenhaus, 1998). Importantly, some of these transient fixations are initiated before the end of the referent's name, demonstrating people's ability to anticipate which word(s) they may be hearing based on partial information. The speed with which extremely brief portions of speech, such as the fragment of a vowel that anticipates the following consonant, begin to influence interpretation is truly remarkable (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; see Fig. 2).

Although these recent studies have shown rapid interpretation and fine sensitivity to phonetic information in the speech signal, they have also documented aspects of the interpretation process that call into question a strictly sequential analysis of speech. First, as illustrated in Fig. 1b, transient fixations to a competitor object with a name that was consistent with the early sounds of the spoken word continue to be observed well after the speech has ceased to match the competitor's name, especially if the competitor is a more common word in the language (e.g., Dahan & Gaskell, 2007; Dahan, Magnuson, & Tanenhaus, 2001). Thus, new information in speech does not result in a swift exclusion of some candidate interpretations; instead, information accumulated over time causes a gradual shift of interpretation in favor of some hypotheses and against others. Second, hearing the pronunciation of a word ("carrot") also elicits the consideration of a rhyming alternative ("parrot"), as reflected in listeners' transient fixations to the picture of the rhyming word (Allopenna et al., 1998). This result cannot be attributed to the occasional misperception of the initial sound because in all cases, people eventually selected the correct referent. Thus, the rhyming portion of the spoken word "carrot" may affect the interpretation of its initial sound: Everything else being equal, the initial sound of "carrot" is more likely to be "p" (due to the influence of *parrot*) than the initial of sound of "candle" is (because there is no such word as *pandle*). This is, in effect, evidence that the later portion of a spoken stimulus can affect the interpretation of an earlier portion. Such "right context" effects have been known for some time, but a mounting body of evidence speaks to the generality and importance of this phenomenon in language comprehension.

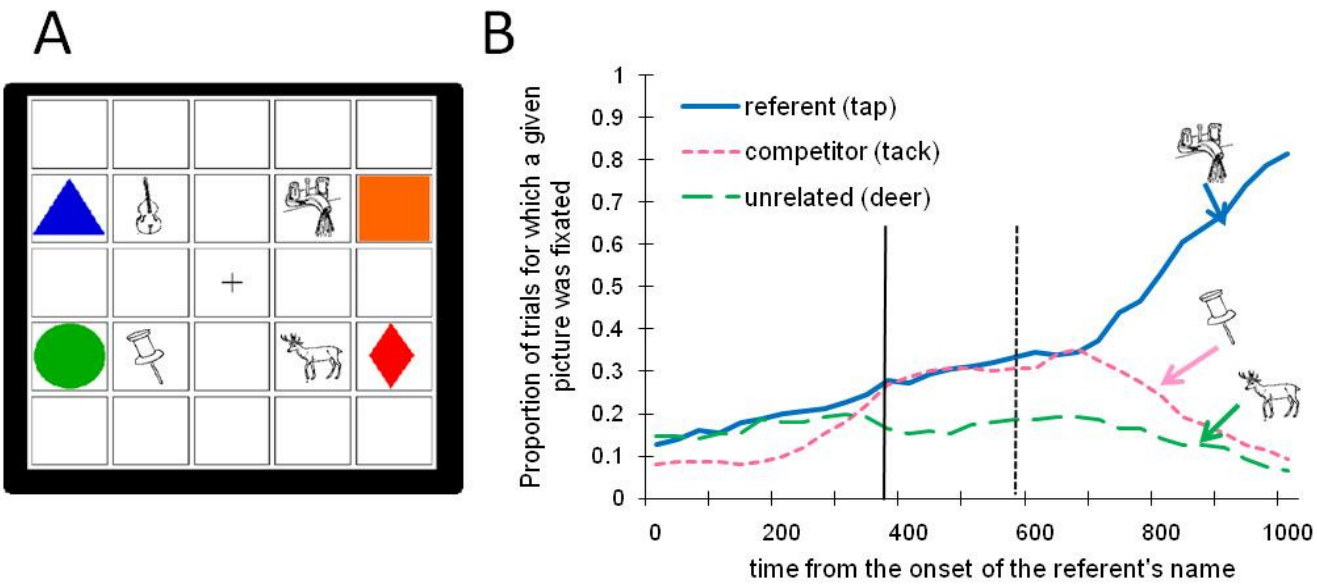

Fig. 2. Experiment showing the speed with which an extremely brief portion of speech—in this case the fragment of a vowel that anticipates the following consonant—begins to influence interpretation. Dahan, Magnuson, Tanenhaus, and Hogan (2001) created two versions of the same word, e.g., *tap*, by concatenating the initial fragment of the same word, *tap*, or of a different word, *tack*, up to and including the vowel, with the final consonant of the word *tap*. The diagram (A) shows the spectral cues in the vowel "a" that anticipate the identity of the following consonant (either "p" or "ck"). The trajectories of each of the three formants (high-amplitude frequencies) in the vowel from *tap* (in black) and *tack* (in red) are superimposed, revealing how the trajectory of the second formant begins to differ about half way through the vowel as a function of the identity of the following consonant. The graph (B) shows the proportion of trials with a fixation to the competitor picture (e.g., a tack) over time, as a function of which version of the referent word participants heard (e.g., "ta[p]p," in black or "ta[ck]p," in red). Participants were much more likely to erroneously fixate on the competitor picture (e.g., the picture of a tack) when hearing a version of "tap" with a vowel that contained spectral cues consistent with an upcoming "ck" than when hearing a version of tap with a vowel that anticipated a "p," even though the duration of the fragment that differed between the two versions is on the order of 100 milliseconds.

When the Present Influences the Past

Evidence for a right-context analysis was provided by the manipulation of fine-grained phonetic properties. For instance, using a recording of the sentence "Did anyone see the gray ship," Repp, Liberman, Eccardt, and Pesetsky (1978) could change the percept of "gray" into "great" by increasing the duration of the initial sound "sh" in "ship." But right-to-left analysis can span a larger window: A study by Ganong (1980) and many studies that followed showed that, when presented with a word that begins with an ambiguous sound, such as the fragment "…ype" preceded by a sound intermediate between "t" and "d," people tend to assign the ambiguous sound the interpretation that, in conjunction with the rest of the stimulus, forms a word. Because this tendency is found only when the initial sound has been artificially altered to fall between the two sound categories, it is unlikely to reflect an intentional bias to give a real-word response. This finding indicates that the rest of the word can influence the interpretation of an earlier ambiguous sound (see also Warren, 1970). Influence of the subsequent context can also be observed when the portion that affects the interpretation of the ambiguous sound is substantially delayed in time—by a few syllables (McMurray, Tanenhaus, & Aslin, 2009) or even a few words. The interpretation of an ambiguous string like "tent" or "dent" is influenced by the sentence context that follows it (e.g., "the t/dent in the forest/fender"; Connine, Blasko, & Hall, 1991). Retroactive effects of this sort can be found even with clearly articulated, undistorted speech. Listeners hearing a real word like "college" or a variant like "gollege" are both a semantically related word like "student," relative to a baseline, showing

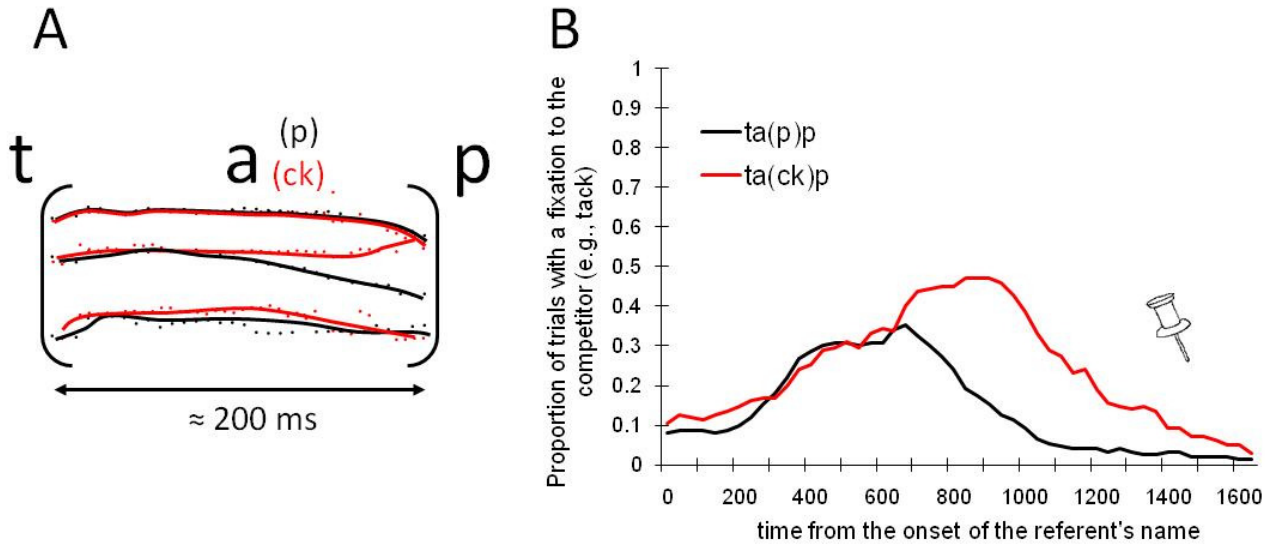

Fig. 2. Experiment showing the speed with which an extremely brief portion of speech—in this case the fragment of a vowel that anticipates the following consonant—begins to influence interpretation. Dahan, Magnuson, Tanenhaus, and Hogan (2001) created two versions of the same word, e.g., *tap*, by concatenating the initial fragment of the same word, *tap*, or of a different word, *tack*, up to and including the vowel, with the final consonant of the word *tap*. The diagram (A) shows the spectral cues in the vowel "a" that anticipate the identity of the following consonant (either "p" or "ck"). The trajectories of each of the three formants (high-amplitude frequencies) in the vowel from *tap* (in black) and *tack* (in red) are superimposed, revealing how the trajectory of the second formant begins to differ about half way through the vowel as a function of the identity of the following consonant. The graph (B) shows the proportion of trials with a fixation to the competitor picture (e.g., a tack) over time, as a function of which version of the referent word participants heard (e.g., "ta[p]p," in black or "ta[ck]p," in red). Participants were much more likely to erroneously fixate on the competitor picture (e.g., the picture of a tack) when hearing a version of "tap" with a vowel that contained spectral cues consistent with an upcoming "ck" than when hearing a version of tap with a vowel that anticipated a "p," even though the duration of the fragment that differed between the two versions is on the order of 100 milliseconds.

that "gollege" triggered the consideration of its real-word counterpart (Connine, Blasko, & Titone, 1993).

The impact of later-arriving information in word recognition is not a marginal, "laboratory" phenomenon; it is a pervasive, fundamental fact about how we interpret spoken language. When assessing listeners' ability to identify words from a recording of an unscripted, spontaneous dialogue, Bard, Shillcock, and Altmann (1988) reported that as many as 20% of the words were accurately recognized only after one or more subsequent words had been heard. Because these subsequent words were often correctly identified themselves, the fact that the recognition of many words was delayed cannot be explained by a mere lag between the arrival of auditory information and its utilization. Subsequent information actively contributes to the interpretation of earlier portions of the speech signal.

Relaxing the requirement of a strictly left-to-right analysis of speech to accommodate right-context effects has far-reaching consequences. In a strictly sequential view of speech comprehension, the listener's decision space becomes smaller over time as the set of possibilities is incrementally reduced. In this view, reconsidering alternative interpretations that had been disfavored early in the sentence but that become likely later on would require revision or backtracking. This revision would be expected to be cognitively costly and to require some time to accomplish. But recent work has failed to find any evidence of such a cost (Dahan &Tanenhaus, 2004; see Fig. 3). The view that listeners can flexibly integrate information, even when this information is improbable or unpredicted, is at odds with the view that their decision space is incrementally restricted as new pieces of information are added. The

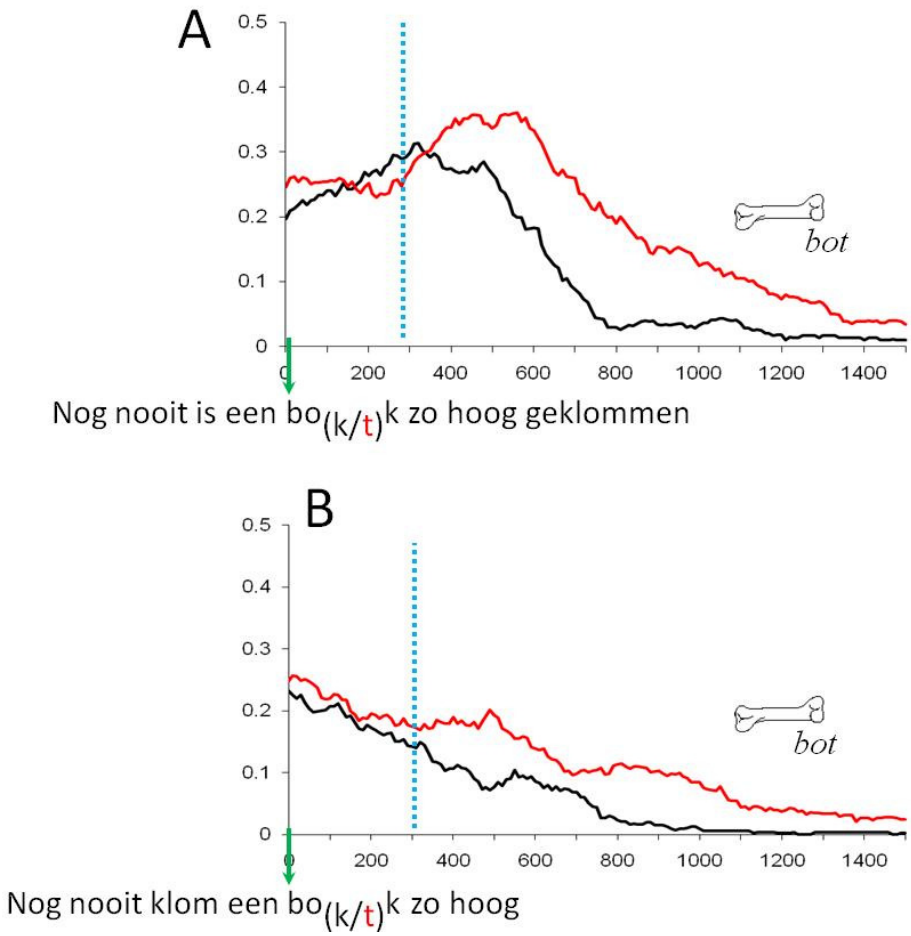

Fig. 3. Results of Dahan and Tanenhaus (2004). Dutch participants heard a sentence while seeing four objects pictured on a display; their task was to select the object mentioned in the accompanying sentence while their eye movements to the objects were monitored. The top graph (a) shows change over time in the proportions of trials with a fixation to the competitor picture (the picture of a bone, bot in Dutch) from the onset of the referent's name (indicated by the green arrow) when the spoken word ''bok'' (goat) contained spectral cues that anticipate the consonant ''k'' (bo[k]k, in black) versus when the spectral cues anticipated the consonant ''t'' (bo[t]k, in red) after a nonconstraining sentential context (''Nog nooit is een bok zo hoog geklommen,'' literally translated Never before had a goat so high climbed). The spectral cues in the spoken word's vowel temporarily modulated interpretation, with more fixations to the bone (bot) when the vowel contained cues that predicted an upcoming ''t'' (red) than when it contained cues that predicted an upcoming ''k'' (black). The point in time at which spectral cues in the vowel began to affect fixations is marked by the vertical blue dotted line. The bottom graph (b) shows change over time in the proportions of trials with a fixation to the competitor picture (a bone, bot in Dutch) from the onset of the referent's name (indicated by the green arrow) when the spoken word ''bok'' (goat in Dutch) contained spectral cues that anticipated the consonant ''k'' as in bok (bo[k]k, in black) versus when the spectral cues anticipated the consonant ''t'' as in ''bot'' (bone in Dutch; bo[t]k, in red) after a constraining sentential context (''Nog nooit klom een bok zo hoog,'' literally translated Never before climbed a goat so high): The verb ''climb'' constrains its subject to be an entity that can climb, which a goat, but not a bone, is. As the spoken word ''bok'' (goat) begins, the probability of fixating on the competitor ''bot'' (bone) decreases, revealing the influence of the verb in rejecting nonclimbing candidates. However, the spectral cues in the vowel mitigates this effect, as shown by the difference between fixations to the bone (bot) when the vowel contains cues that predict an upcoming ''t'' (red) than when it contains cues that predict an upcoming ''k'' (black). The point in time at which spectral cues in the vowel began to affect interpretation, marked by the vertical blue dotted line, is quite similar to what was observed after a nonconstraining context. Thus, spectral cues in vowels can readily modulate word interpretation even when the context and the initial portion of the spoken word, in conjunction with the visual display, have converged toward a unique interpretation.

interpretation of a stretch of speech appears to result from the integration of a number of constraints over a larger temporal window than typically assumed under the sequential-analysis view. A similar conclusion can be drawn from recent evidence that people retrieve the meaning of a monosyllabic word (e.g., "pain") upon hearing a longer word for which it forms the last syllable (e.g., "champagne") if the semantic context preceding the longer word was more congruent with the short than the long word (e.g., "The patient asked the nurse when the champagne would be cold enough to be served"; van Alphen & van Berkum, in press). A strictly left-to-right view is ill-equipped to account for such phenomena.

Word Recognition as a Perceptual Choice

Early research on real-time speech comprehension put special emphasis on the sequential property of speech by assuming that analysis and decision making in speech interpretation proceed at the same pace with which information in the speech signal arrives to the senses. The picture that is emerging from more recent work suggests a more complex process, one in which percepts during speech comprehension emerge from both anticipation of upcoming information and integration over a larger temporal window.

Renouncing the view of a strictly linear analysis of the speech signal contributes to bringing theories of spoken-language comprehension in line with what is known about the details of speech perception. Individual speech sounds like consonants and vowels do not form separate, discrete events like letters on the page; they overlap substantially and are distributed over time. In addition, interpretation of one sound depends on interpretation of nearby sounds, in both directions. The temporal grain at which decisions about the words of an utterance are made cannot be as fine as early theories posited.

Theories of perception for which choice is the result of a dynamical process also point to the inadequacies of the strictly sequential analysis of speech. Such theories assume that perceptual choice is a process that operates over intrinsically noisy perceptual information that accumulates over time, even for static stimuli (e.g., Townsend & Ashby, 1983). Accordingly, decisions regarding the identity of spoken words are not typically driven by a single sample of information from a limited snippet of speech, because the neural encoding of speech is too noisy. As more neural samples of the same snippet accumulate over time, however, the distribution over the samples converges toward a stable interpretation. Importantly, because speech also changes over time, the integration of samples is likely to encompass a larger temporal window than the strictly left-to-right view envisions. This is not to say that the perceptual system fails to make use of early-arriving phonetic information, which does provide partial cues to the words being said. This is why numerous studies, including those relying on eye movements to assess the uptake on auditory information, have found evidence that the earliest moments influence choices. However, we argue, the ultimate (i.e., asymptotic) decision is reached after more time has passed and more information has accumulated.

As speech-comprehension theories move toward allowing for integration over a larger temporal window, the role played by sensory or working memory becomes more apparent: Words may not always be recognized in the order they were spoken, but interpretation relies on keeping words in the intended sequence. Some computational models have explicitly incorporated mechanisms that allow for retention of the true word sequence (Grossberg & Myers, 2000). In fact, the most enduring and perhaps most successful model of spoken-word recognition, the TRACE model (McClelland & Elman, 1986), has an architecture that accommodates the empirical data reported here, because it explicitly distinguishes the timeline of the unfolding of the phonetic signal from the timeline imposed by the continuous integration of information leading to the emergence of stable percepts of words or sentences. Nonetheless, the model's ability to simulate human performance, such as the data reported here, ultimately hinges on relatively unexplored aspects of the model, such as the dynamics with which stable percepts of words emerge out of the noisy representation of a

continuously changing signal.

More generally, the study of speech comprehension offers a unique opportunity to examine how people evaluate noisy and rapidly changing perceptual information, and may ultimately provide important constraints to current theories of perceptual choice.

Declaration of Conflicting Interests

The author declared that she had no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by research grants from the National Science Foundation under Grant No. 0433567 and the National Institutes of Health (R01 HD 049742-1).

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.

Bard, E.G., Shillcock, R.C., & Altmann, G.T.M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics*, *44*, 395–408.

Cole, R.A. (1981). Perception of fluent speech by children and adults. In H. Winitz (Ed.), *Native language and foreign language acquisition* (pp. 92–109). New York: The New York Academy of Sciences.

Connine, C.M., Blasko, D.G., & Hall, M. (1991). Effects of subsequence sentence context in auditory word recognition: temporal and linguistic constraints. *Journal of Memory and Language*, *30*, 234–250.

Connine, C.M., Blasko, D.G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, *32*, 193–210.

Dahan, D., & Gaskell, M.G. (2007). Temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. *Journal of Memory and Language*, *57*, 483–501.

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.

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

Dahan, D., & Tanenhaus, M.K. (2004). Continous 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.

Ganong, W.F., III (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*, 267–283.

Grossberg, S., & Myers, C.W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review*, *107*, 735–767.

Marslen-Wilson, W.D. (1975). Sentence perception as an interactive parallel process. *Science*, *189*, 226–228.

Marslen-Wilson, W.D., & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1–71.

McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McMurray, B., Tanenhaus, M.K., & Aslin, R.N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition.

Journal of Memory and Language, *60*, 65–91.

Repp, B.H., Liberman, A.M., Eccardt, T., & Pesetsky, D. (1978) Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, *4*, 621–637.

Townsend, J.T., & Ashby, F.G. (1983). *The stochastic modeling of elementary psychological processes*. New York: Cambridge University Press.

van Alphen, P.M., & van Berkum, J.J.A. (in press). Is there pain in champagne? Semantic involvement of words within words during sense-making. *Journal of Cognitive Neuroscience*.

Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392–393.