Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 57 (2007) 483–501
Journal of Memory and Language
www.elsevier.com/locate/jml

The temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition

Delphine Dahan [a,*], M. Gareth Gaskell [b]

[a] Psychology Department, University of Pennsylvania, 3401 Walnut Street, Room 302C, Philadelphia, PA 19104-6228, USA
[b] Department of Psychology, University of York, York, YO10 5DD, UK

Received 30 June 2006; revision received 8 January 2007
Available online 20 February 2007

Abstract

Two experiments examined the dynamics of lexical activation in spoken-word recognition. In both, the key materials were pairs of onset-matched picturable nouns varying in frequency. Pictures associated with these words, plus two distractor pictures were displayed. A gating task, in which participants identified the picture associated with gradually lengthening fragments of spoken words, examined the availability of discriminating cues in the speech waveforms for these pairs. There was a clear frequency bias in participants' responses to short, ambiguous fragments, followed by a temporal window in which discriminating information gradually became available. A visual-world experiment examined speech contingent eye movements. Fixation analyses suggested that frequency influences lexical competition well beyond the point in the speech signal at which the spoken word has been fully discriminated from its competitor (as identified using gating). Taken together, these data support models in which the processing dynamics of lexical activation are a limiting factor on recognition speed, over and above the temporal unfolding of the speech signal.

Keywords: Spoken-word recognition; Eye-tracking; Gating; Dynamical processing; Lexical competition

Comprehending an utterance requires the identification of its component words. How this process is conceived has been strongly influenced by the seminal work of Marslen-Wilson and colleagues (Marslen-Wilson, 1973, 1975; Marslen-Wilson & Welsh, 1978). In a marked departure from the then dominant views, Marslen-Wilson argued that the process of recognizing the presence of a given word in the spoken input depends not just on the amount of evidence in favor of that word, but also on the amount of evidence in favor of alternatives. The recognition process is thus contingent in nature. Furthermore, spoken words are acoustic patterns that become available to the listener gradually, and the sequentiality with which sensory information is received plays a critical role. The point in time at which a spoken word can be recognized, argued Marslen-Wilson, can be precisely identified as the point where the input has excluded all candidates but one (i.e., the word's "uniqueness" point). This point can be estimated from phonemic transcriptions and corresponds to the phoneme position at which only one word remains compatible with the phonemic sequence, from left to right. Empirical results provided strong support to this view. When measuring lexical-decision response times to spoken stimuli from their acoustic onset, Marslen-Wilson (1984) reported

* Corresponding author. Fax: +1 215 573 9247.
E-mail address: dahan@psych.upenn.edu (D. Dahan).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.01.001

a linear relationship between a nonword response latency and the position of the phoneme at which the sequence diverged from all existing words. This led Marslen-Wilson to conclude that "human speech processing does permit optimal real-time information extraction" (Marslen-Wilson, 1984) [p. 141]. Although Marslen-Wilson (1987) revised his theory to allow phonetic input to provide gradient, rather than all-or-none, support to word candidates (embodied in the metaphor of gradually-changing word activation), the assumption that word recognition makes immediate use of the signal as it unfolds has remained. The present study revisits these claims by examining the temporal dynamics of spoken-word recognition. This issue is important because it lies at the heart of language processing in general: At all levels we need to understand the linkage between gradual changes in the information contained in the signal (the speech waveform) and the cognitive impact of these changes.

The advent of the visual-world paradigm as a method of estimating lexical activation over time has provided researchers with a potentially powerful tool to study the time course of spoken-word recognition (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In many studies using this method, a display composed of a few pictured objects is presented, followed by a spoken utterance that mentions the name of one of the objects. Participants are instructed to click on the named object using the computer mouse. The location of their gaze is monitored as the speech is heard and until participants have clicked on the object. When considered over many trials, gaze location can be transformed into probabilities of fixating each of the objects over time, starting at the onset of the spoken word. One consistent finding emerging from this research is the immediacy of the uptake of acoustic information. Allowing for a well-established 200-ms delay for the programming and launching of an eye movement (Hallett, 1986), the probability of fixating on a picture with a name that matches the spoken input begins to increase immediately after the input becomes available (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, & Tanenhaus, 2001). For instance, if the display contains the pictures of a candle, a candy, a pear, and a necklace, the probability of fixating on the candle or the candy begins to increase about 200 ms after the onset of the spoken word *candle*, concurrent with a decrease in the probability of fixating on the pear or the necklace. Note that the time locking between gaze location and lexical processing of the spoken input requires that participants have accumulated enough information about the visual display to be able to direct their attention and gaze to linguistically relevant locations. How this knowledge accumulates in the course of a trial to influence fixations has received little attention.

Because fixation probabilities appear closely time-locked to the input and vary continuously over time, Allopenna et al. (1998) took the step of developing a linking hypothesis between fixation probabilities and lexical activation. Based on lexical activation generated by an implemented model of spoken-word recognition, the TRACE model (McClelland & Elman, 1986), Allopenna et al. found a very close fit between the behavioral data and underlying activation functions, given simple assumptions about the task. In particular, they assumed that the probability of fixating on a given picture at time t reflects the strength of evidence that the picture is the referent relative to the strength of evidence supporting the other alternatives on the display. As alluded to earlier, this mapping between lexical activation and eye movements may be less straightforward if people are still gathering information about the visual alternatives on the display as the spoken input unfolds.

The Allopenna et al. study is an important first step in establishing the validity of the visual-world paradigm as a tool to study the temporal dynamics of lexical processing. The current study is aimed at extending this effort, focusing on the time course of competitor's activation. In much of the published research examining lexical processing with the visual-world paradigm, the probability of fixating a competitor picture (e.g., a candy when hearing *candle*) remains higher than the probability of fixating a distractor picture (e.g., a necklace) for a substantial amount of time after target and competitor fixation probabilities have started diverging, an indication that the input has started providing more support for the target than for the competitor. This raises the possibility that the impact of the phonetic input on lexical activation has an immediate onset but evolves only gradually over time. A gradual change could be attributed to one or both of two factors. First, the speech signal itself may provide probabilistic information whose diagnostic value changes over time. Indeed, the articulatory gestures involved in the production of successive speech segments are dynamical events whose acoustic consequences and their informational value may change over time. In addition, the perceptual process may be internally noisy, and its outcome stochastic, rather than deterministic; if so, the accumulation of evidence over time, even in the absence of any new information in the signal, would increase the signal-to-noise ratio.

While the first factor is often acknowledged, and sometimes approximated, in models of spoken-word recognition, the notions of stochastic perceptual and/or decisional processes and processing-internal dynamics have not been widely embraced, despite being central to major theories of perceptual choice (e.g., Massaro, 1998; Ratcliff, 1978; Ratcliff, Van Zandt, & McKoon, 1999; Usher & McClelland, 2001) and, more recently, of lexical processing (Norris, 2006; Wagenmakers et al., 2004). Norris (2006) developed a Bayesian model

of visual-word recognition in which sensory information is noisy and sequentially sampled, despite being physically instantaneously available to the reader. Early in the process, the amount of extracted information is limited and the signal-to-noise ratio low. However, as time increases, the number of samples from the input accumulates, and extracted information tends to converge on the correct interpretation. Thus, ambiguity in interpretation declines over time, even though the physical stimulus itself does not change.

The notion of internal dynamics is also central to connectionist models that incorporate some form of recurrence. In such models, lexical activation (whether embodied among localist or distributed representations) changes over time, even without any changes in the sensory input or in the information extracted from the sensory input, until a stable state has been reached. The dynamics depend on the complex interaction between the information in the sensory input and the state space, i.e., the distance between basins of attraction (which correspond to categorically incompatible representations, e.g., words) and their depth (see Kawamoto, 1993; for a good illustration of such models).

The goal of the present study is to provide evidence that the recognition of a spoken word and the rejection of its competitors is a gradual process, constrained both by the dynamics with which acoustic cues to phonetic features change over time and by the internal dynamics of the perceptual process itself.

There is a general consensus that lexical activation associated with a word candidate increases with increasing phonetic support from the input. Furthermore, given that frequent words are recognized more rapidly than rarer words (and more accurately in degraded conditions), most models assume that frequency of occurrence amplifies the impact of the phonetic match, with the activation of high-frequency candidates having a higher resting level or increasing faster than that of low-frequency candidates. This fits a Bayesian framework, where an optimal decision process integrates the degree of similarity between the input and lexical candidates with prior biases (Norris, 2006). Conversely, a lack of support from the phonetic input decreases a word candidate's activation. Two mechanisms for activation reduction have been proposed. Models incorporating bottom-up inhibition employ a specific link between the speech input that conflicts with a particular word and the representation of that word (e.g., Norris, 1994), allowing the negative evidence to impact directly on the activation of the word. Alternatively or in addition, the activation of mismatching candidates may be decreased via a more indirect route in which conflicting evidence provides positive support for a neighboring lexical candidate, which then reduces the activation of the mismatching candidate through lateral inhibition at the lexical level (as in the TRACE model or the Shortlist model (Norris, 1994)). In cases where the competition environment is sparse, these two mechanisms can provide quite different predictions as to the effects of mismatch, but in most cases, both provide an adequate means of reducing activations.

In currently implemented models, where sensory input is assumed to become available to the processing system as it unfolds over time, the temporal dynamics associated with bottom-up inhibition alone on the activation of a mismatching candidate should reflect the timing and strength with which cues in the spoken input disfavor this candidate. However, for models assuming sequential sampling of noisy input and/or those incorporating lateral inhibition, the dynamics should strongly be constrained by the assumed internal dynamics of the model, captured by the number of processing cycles a model goes through before receiving the next input slice. As illustrated by simple simulations conducted with the TRACE model (using jTRACE, a user-friendly re-implementation of TRACE, see Strauss, Harris, and Magnuson, in press), an increase in the number of processing cycles per input time slice affects the time course of candidate activation, thereby reducing the temporal window over which a difference in activation between a high-frequency candidate and a low-frequency candidate is observable (Fig. 1). Fast internal processing in effect amplifies potentially small differences in the bottom-up support that various candidates receive. As we will show, the dynamics of lexical activation over time are most consistent with a position where the impact of discriminating information in the signal is gradually amplified over time, and in which the impact of initial biases gradually decreases during and beyond the presentation of a spoken word.

The dynamics of spoken-word recognition have often been examined using the gating task (Grosjean, 1980). In this paradigm, increasingly large portions of a spoken word are presented to participants. After each portion (or gate), participants are asked to decide which word they are hearing and to provide an estimate of their confidence in that decision. For a given word, the *isolation point* is defined as the duration of the gate at which the participant correctly identified the stimulus word and did not subsequently change his/her guess. Grosjean (1980) found earlier isolation points for high-frequency than low-frequency words. Furthermore, errors tend to be words of higher frequency than the targets (see also Tyler, 1984).

Marslen-Wilson (1990) found earlier isolation points for words with low-frequency competitors than for words with high-frequency competitors when the frequency of the gated words themselves was matched. Importantly, the competitor-frequency difference was not found on *recognition points* (defined as the point where the listener first gets the word right and has at least an 80% confidence in his/her choice). Thus, the

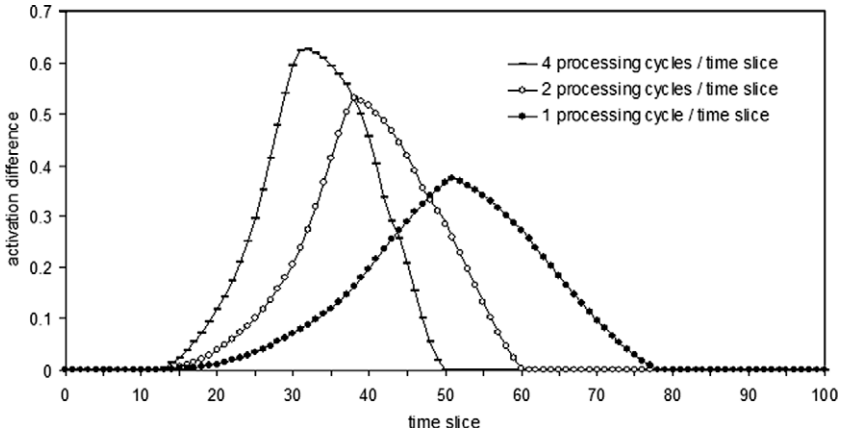

Fig. 1. jTRACE simulations. Difference over time slice in the activation of a high-frequency competitor (i.e., the word "dark" when the spoken input is *dart*) and a low-frequency competitor (i.e., the word "dart" when the spoken input is *dark*), as a function of the number of processing cycles per input time slice. *Note*. Simulations were run with jTRACE, version a.47, using the default lexicon. All parameters were set to their default values, except for the phoneme to word connection weights (frq $p \rightarrow w$ wts), which was set to 0.13, the value used in the Dahan et al. (2001) simulations of frequency effects. The parameter nrep, which controls the number of processing cycles per time slice, was set to 1, −2, or −4.

prevalence of high-frequency responses may simply reflect response biases only at points in the spoken input where the sensory information is still ambiguous. A similar conclusion was reported by Zwitserlood (1989). Using a cross-modal semantic priming technique, she observed a non-significant trend toward more priming for a high-frequency than a low-frequency candidate when both were compatible with the incomplete spoken prime. This trend disappeared when the duration of the spoken prime increased. Likewise, Marslen-Wilson (1990) failed to find evidence of the impact of competitor frequency on the recognition of a spoken word when participants performed a lexical-decision or a naming task, where responses are generally made after the entire word has been heard. This led Marslen-Wilson to conclude that competitor frequency affects word recognition, but has "a transient effect that has dissipated by the time the end of the word is reached" (Marslen-Wilson, 1990) [p. 158]. From this perspective, the timing of the recognition process is mostly affected by the arrival of sensory input, i.e., the external sequentiality of spoken words. The state of lexical processing reached at time t is determined, to a large degree, by the content of the physical signal available at that point.

This conclusion stands in stark contrast with a large body of research that has demonstrated the impact of the number and frequency of competitors on the ease and speed of recognition of a spoken word (e.g., Luce, 1986; Luce & Large, 2001; Vitevitch & Luce, 1998). For example, Luce (1986; see also Luce & Pisoni 1998) demonstrated the effect of the number and/or frequency of similar-sounding words on the recognition of a given spoken word. This effect was found on latencies to shadow a spoken word or to make speeded lexical decision. These effects (and related ones) suggest that the impact of competition from alternative words can extend beyond the acoustic offset of a spoken word.

A study by Zwitserlood and Schriefers (1995) is also relevant here. The size of a priming effect caused by the presentation of the initial portion of a spoken word on the processing of a semantically related target increased when a short (100 ms) delay was added between prime offset and target onset. They concluded that processing time alone, without any change in the sensory input, can influence lexical activation.

Thus, it appears that two incompatible views coexist in current accounts of spoken-word recognition. One view holds that word recognition is closely time locked to the arrival of disambiguating information in the signal. Biases, such as frequency of occurrence, operate only while the spoken input is insufficient for discriminating between lexical candidates. Once the signal begins to support some candidates more than others, the perceptual system makes immediate use of the input and inconsistent candidates are expeditiously discarded from consideration, with preexisting biases swiftly discounted. The alternative view assumes that the processing of the speech signal is inherently stochastic and dynamical. This view holds that biasing factors such as frequency effects are influential over a longer timescale, and do not dissipate as soon as discriminating information in the speech signal is encountered.

The co-existence of these two views is most apparent in a set of studies where listeners were exposed to spoken words containing subphonemic mismatches (cf. Whalen, 1984). Such stimuli are created by cross-splicing two

monosyllabic stimuli that differ in their final consonant, e.g., *neck* and *net*. A cross-spliced version of *net* may consist of the final /t/ of a token of *net* and the initial consonant and vowel of another word, *neck*, or of a nonword, *nep*. The cross-spliced stimulus contains coarticulatory cues in the vowel that anticipate an upcoming /k/ or /p/, even though a /t/ actually follows. An identity-spliced version of *net*, created by splicing together two different tokens of *net*, contains no subphonemic mismatch. When listeners were asked to make a lexical decision to spoken words, they responded more slowly to cross-spliced versions than to identity-spliced counterparts. Importantly, whether the initial portion of the cross-spliced word originated from a different word or from a nonword did not measurably affect performance (Marslen-Wilson & Warren, 1994; see also McQueen, Norris, & Cutler, 1999). This result was interpreted by Norris, McQueen, and Cutler (2000) as evidence that, during the processing of a word cross-spliced with another existing word (e.g., $ne_{(ck)}t$), transient competition between the word supported by coarticulatory cues in the vowel (e.g., *neck*) and the word supported by the burst from the final stop consonant (e.g., *net*) is resolved so quickly that lexical-decision latencies fail to reveal it. This, in turn, was taken as evidence supporting fast internal processing dynamics, as implemented in Shortlist. However, Dahan, Magnuson, Tanenhaus, and Hogan (2001) conducted a visual-world version of the study, where people clicked on a named referent (e.g., a net). On critical trials, the referent's name had been spliced with an existing word or a nonword. The probability of fixating on the referent picture over time increased more slowly when the initial portion of the spliced word originated from an existing word (e.g., $ne_{(ck)}t$), compared to when it originated from a nonword (e.g., $ne_{(p)}t$). Furthermore, this effect extended substantially beyond the acoustic offset of the stimulus. This result is inconsistent with the idea that the information in the final portion of the cross-spliced word rapidly overcomes the advantage that the initial portion had given to the competitor. Rather, the extended competition effect observed in the visual-world study suggests a temporally gradual impact of the final sound on the ultimate interpretation of the spoken input, which may be attributed to the sequential sampling of noisy input and/or to stochastic and dynamical processing of the input. However, it is conceivable that the extended competition effects observed in this study were a special consequence of the cross-splicing manipulation, in which listeners responded to words containing conflicting phonetic cues.

The present study examines the dynamics of lexical activation during the perception of unaltered speech, using two types of behavioral data: gating responses and eye movements in the visual-world paradigm. In both tasks, participants saw a computer display with four pictured objects. In the gating task, they heard increasingly long fragments of a spoken word and were asked to indicate, after each gate, which of the four objects was being named and their confidence in their choice. In the visual-world task, participants heard the name of one of the four objects in isolation and were instructed to click on the referent with the computer mouse. Their eye gaze was monitored as the spoken stimulus was heard and until they clicked on the referent object.

These two tasks provide an estimate of target and competitor consideration over time. However, they differ fundamentally in terms of their dynamics of processing. Gating responses can be generated after substantial processing time has passed after the presentation of each gate, and most likely reflect lexical interpretation given the available sensory input after internal processing has reached an asymptotic, stable state. Thus, gating can be described as an asymptotic choice paradigm (McClelland, 1993). Gating responses over time should closely reflect the unfolding of the spoken input and the information it conveys in distinguishing target and competitor words. Eye movements, on the other hand, occur concurrently with the spoken input in real time. Thus, fixations initiated at time t should reflect the processing state reached at that time, itself determined by both the available input and internal processing dynamics. Thus, differences in gating and eye-movement responses may be taken to reflect the impact of the internal processing dynamics on lexical interpretation.

Analyses of gating and eye-tracking data will proceed in the following way. First, we report on responses from the gating task. Here, we show that, even when responses are generated after processing has reached a stable state, identifying a spoken word and distinguishing it from its competitor(s) cannot be linked to a particular piece of sensory information in the input. Instead, we can identify a disambiguating period, during which sensory information accumulates and gradually changes the balance of evidence for each lexical alternative. This suggests that the speech signal itself provides probabilistic information whose value changes over time. We then use the gating responses to determine the point in the auditory stimulus at which listeners are essentially unanimous in correctly identifying the target word. Examination of eye-tracking responses relative to the point in time reveal that even when the speech signal contains sufficient information for ruling out a lexical competitor, listeners consider this competitor to a greater degree than they consider phonologically unrelated distractors. In addition, competitor activation remains greater for high-frequency words than for low-frequency words. These findings lead us to conclude that the sensory input and/or the perceptual processes underlying the recognition of spoken words are noisy and stochastic, and subject to internal dynamics, a fact that makes the mapping of sensory input onto perceptual choices more complex than is often assumed.

A supplementary goal of this study is to contribute to a better understanding of how eye movements over the course of a trial reflect lexical activation. As pointed out earlier, signal-driven fixations to pictured referents rely on a complex interaction between lexical interpretation of the spoken input and the information that participants have gathered and/or are gathering about the visual display. This aspect is seldom acknowledged and generally poorly understood. However, as the visual-world paradigm is increasingly used to test fine-grained hypotheses on lexical processing, a better understanding of the process by which participants orient their gaze to spatial locations associated with pictured objects in the course of a trial is becoming essential. As a step in this direction, we report here a series of analyses on the eye-movement data that take into account the relationship between having attended and fixated a picture and refixating it later in the trial.

Method

Participants

Participants were 60 college students from the University of Nijmegen, the Netherlands. Half of them took part in the gating version of the experiment, the other half, in the eye-tracking version.

Materials

Twenty-eight pairs of picturable Dutch nouns overlapping at onset were selected. Within each pair, one of the nouns had a high frequency of occurrence, and the other, a low frequency (e.g., *koffie* [coffee] and *koffer* [suitcase]). Based on the CELEX database (Baayen, Piepenbrock, & van Rijn, 1993), the high-frequency items had an average log frequency per million of 1.7 ($\sigma = 0.6$), compared to 0.8 for the low-frequency items ($\sigma = 0.5$). The two members of each pair were matched for their number of syllables and the syllabic structure of their overlapping portion. In the interest of maximizing the frequency difference between the two members of each pair while using only picturable words, however, the extent of the phonemic overlap between the two items could not be equated across pairs. The Appendix A lists the pairs, the number of phonemes they share at onset, and the frequency of each member. In order to form a four-item display, two additional phonologically unrelated picturable nouns were associated with each onset-overlapping pair (e.g., *hond* [dog] and *spiegel* [mirror]). Each noun was matched for frequency with one item of the pair. The high-frequency matched distractors had an average log frequency of 1.7 ($\sigma = 0.5$), and the low-frequency matched distractors, of 0.7 ($\sigma = 0.5$).

A black and white line-drawing was selected for each of the words. The extent to which these depicted the corresponding nouns was measured using a picture-naming task, administered to an independent group of 15 Dutch speakers, selected from the same population as that of the main experiment. Naming responses to each item of an experimental pair were coded to evaluate the identification of the pictured object (i.e., whether the concept depicted was accurately identified, irrespective of whether the label used to describe the concept was the intended one or a synonym) and the use of its intended specific name. High- and low-frequency pictures depicted their respective concepts quite accurately (respectively, 95 and 94% correct picture identification, with 9 pairs where the identification was more accurate for the high-frequency than the low-frequency item, and 4 pairs with the reverse tendency) and were generally attributed their precise intended names (88 and 87% correct picture labeling for the high- and low-frequency pictures, with 15 pairs for which the labeling of the high-frequency item was more accurate than that of its low-frequency counterpart, and 8 with the reverse tendency).

In addition to the 28 critical trials, 70 filler trials were constructed to be used in the visual-world version of the study; 35 were composed of four phonologically and semantically unrelated words; the other 35 trials included two onset-overlapping words, neither of which played the role of target during the experiment. Pictures for each word were selected from a large database of pictures that have been used in other visual-world studies, ensuring homogeneity in the pictures' overall appearance between experimental and filler trials.

Spoken words, produced in isolation, were read aloud by a female native speaker of Dutch and recorded on DAT-tape in a sound-proof room. Each spoken word was then digitized and edited. The average duration of the experimental target words was 528 ms (538 ms for high-frequency words, 518 ms for low-frequency words), varying from 385 to 804 ms.

Gated versions of each spoken word were created. In order to obtain a detailed picture of word recognition over time, we opted for a short gate increment (i.e., 20 ms). However, to avoid fatigue effects, only the portion of each word deemed to be ambiguous between the two pair items was gated. Thus, for each word pair, the longest initial portion before disambiguating information could be heard was assessed by a native speaker of Dutch. This first gate, whose duration was matched between the two words of each pair, was located between 50 and 360 ms after word onset. The subsequent gates consisted of increments of 20 ms from the initial gate, until each member of the pair could be clearly identified and distinguished from its counterpart, as determined by the same native speaker of Dutch. To avoid auditory artifacts, the last 2 ms of each gate were faded. The last gate, also matched between the two words

of each pair, was located between 170 and 580 ms after word onset. Thus, each word was incrementally presented over an average of 13 gates (ranging from 8 to 20), as well as in its totality. The eye-tracking experiment used only the complete words.

Our method of matching equated the portion of the high-frequency word and its low-frequency counterpart presented at each gate in terms of absolute duration. As pointed out by an anonymous reviewer, high-frequency words are typically shorter in duration than low-frequency words. This effect is found even when the phonological structure of the words is controlled, as in the case of homophones (Jurafsky, Bell, & Girand, 2002). Although the overall duration of our low- and high-frequency items showed a tendency in the opposite direction, the portion of the word that phonemically overlaps across these items could conceivably have been produced faster for the high-frequency than for the low-frequency item. If so, equating the duration of gates across the high- and low-frequency items of each pair may have resulted in presenting relatively more of the overlapping portion of the high-frequency word than of its low-frequency counterpart. Consequently, based on portions of the spoken word of the same length, a high-frequency word may be discriminated from its low-frequency competitor better than a low-frequency counterpart would be from its high-frequency competitor. Importantly, however, the logic employed in this study avoided this pitfall. As described in the results section, the point in time at which spoken words were (nearly) perfectly discriminated from their competitors was assessed for each pair, taking into account gating responses made to both high-frequency and low-frequency items.

Design and procedure

Gating

Participants were seated at a comfortable distance from a computer screen. On each trial, four pictures, arranged as the four corners of an imaginary square and each associated with a number, appeared on the screen. The upper left picture was numbered one, the upper right picture, two, the lower left picture, three, and the lower right picture, four. Following the conventional gating procedure, a 9-point scale was displayed at the bottom of the screen, with 1 labeled as uncertain (*onzeker*) and 9, certain (*zeker*). Simultaneously with the appearance of the display, the first gate of the target word was played through headphones. Prior to the experiment, participants were instructed that they would hear a truncated spoken word of increasing length. They were to decide, after each stimulus, which picture they were hearing the name of, and give an estimate of their confidence in their choice. They indicated their choice by entering the number associated with the picture of their choice on the keyboard (from 1 to 4), followed by their confidence rating (from 1 to 9). Once both responses had been entered, the next gate was played. The next trial consisted of presenting the same set of pictures, occupying the same positions, and the next gate of the same spoken word, until the complete target word was heard. The subsequent trial showed a new four-picture display and word fragments of increasing length were played. On each new display, the positions of the pictures were randomized. Each participant completed 28 series of trials, where a series corresponds to the same display and of fragments of increasing length of the same word. On half of the 28 series, the spoken word was the high-frequency item of the word pair (i.e., high-frequency condition trials); on the other half, the spoken word was the low-frequency item of the remaining pairs (i.e., low-frequency condition trials). Two counterbalanced lists were constructed, varying which items of the pairs were presented, and ensuring that participants never heard both members of a word pair. Participants were randomly assigned to lists, and for each list, 10 random orders were created.

Eye-tracking

Participants were seated at a comfortable distance from a computer screen. Eye movements were monitored with an SMI Eyelink system, sampling at 250 Hz. The head-mounted eye tracker was first fitted onto the participant's head, and a brief calibration procedure was performed. On each trial, a central fixation point appeared on the screen for 500 ms, followed by a blank screen for 600 ms. Then, a 5×5 grid with four pictures, four geometric shapes, and a central cross appeared on the screen 500 ms before the presentation of the complete spoken word. Prior to the experiment, participants were instructed that they would hear a word referring to one of the pictured objects on the screen. Their task was to click on the picture and move it above or below the geometric shape adjacent to it, using the computer mouse. Positions of the pictures were randomized across four fixed positions of the grid. The positions of the geometric shapes were fixed. The edges of the pictures were approximately 4 cm apart; the distance between the central cross and the closest edge was roughly 3 cm. (1 cm corresponded to approximately 1° of visual arc.) Participants were under no time pressure to perform the action. After the participant moved the picture, the experimenter pressed a button to initiate the next trial. Every five trials, a central fixation point appeared on the screen, allowing for automatic drift correction. Two counterbalanced lists were created, varying which item of each pair was the target and ensuring that each participant heard one member of each pair. Three

trial orders were created for each list; participants were evenly distributed across lists and orders.

Results and discussion

Gating

Each subject provided 373 choices. A small proportion of the choices (0.3%, 30 choices) was discarded because of technical failures (2 responses) or because listeners entered a value outside of the range of possible values (i.e., from 1 to 4), probably from mistyping (28 responses). These trials were excluded from all analyses. Analyses on choices were conducted with no reference to their confidence estimates, as there is no straightforward analog of confidence to use when analyzing fixations in the eye-tracking study.

Overall, participants almost always identified the complete or truncated spoken word as the name of the target or the competitor pictures (11,106 out of 11,160 responses, 99.5%), giving very few distractor responses. When the first gate was presented, the proportions of target responses were 68 and 40% in the high- and low-frequency conditions, respectively; conversely, the proportion of competitor responses were 31 and 60% in the high- and low-frequency conditions, respectively (where the competitor was of low- and high-frequency, respectively). An ANOVA performed on the arcsine transformed proportions of responses revealed a main effect of response frequency (63 and 35% for high-and low-frequency responses, respectively, with a 95% confidence interval of $\pm 6\%$, $F1(1,29) = 93.1$, $p < .0001$, $F2(1,27) = 21.7$, $p < .0001$, $\min F'(1,39) = 17.6$, $p < .001$). This effect reveals a strong bias in giving high-frequency responses, even when participants are selecting their response from a closed set of visible alternatives. There was also a marginal preference for giving target responses over competitor responses (54% vs. 45%, with a 95% confidence interval of effect of $\pm 9\%$, $F1(1,29) = 4.03$, $p = .054$, $F2(1,27) = 4.25$, $p = .049$, $\min F'(1,56) = 2.1$, $p = .15$), suggesting that subtle acoustic cues to word identity were sometimes present even in the speech portion corresponding to the first gate. No interaction between these two factors was found (Fs < 1). When the complete word was presented, it was correctly identified as the name of the target 98% of the time (98.1% for high-frequency targets and 98.6% for the low-frequency targets).

In order to establish when the sensory input provided information allowing the listener to disambiguate target and competitor, we computed, for each word pair and at each gate, the proportion of participants who gave target or competitor responses in the high-frequency and low-frequency condition (i.e., when the spoken input was either the high-frequency or low-frequency item of the pair, respectively). Because the duration of the complete words differed between the high-frequency and the low-frequency item of each pair, the responses made to complete words were not included here. For each gate, we subtracted the competitor proportion from its frequency-matched target proportion for high-frequency and low-frequency conditions separately, and averaged the two differences. The equation for the calculation of the target-competitor proportion difference is given in [1]:

$$[(p(R_{\mathrm{HF}}|W_{\mathrm{HF}}) - p(R_{\mathrm{HF}}|W_{\mathrm{LF}})) + (p(R_{\mathrm{LF}}|W_{\mathrm{LF}}) - p(R_{\mathrm{LF}}|W_{\mathrm{HF}}))]/2 \quad (1)$$

where $p(R_{\mathrm{HF}} \mid W_{\mathrm{HF}})$ corresponds to the proportion of participants, out of the 15 tested, who "correctly" gave a high-frequency (HF) target response (R) to the high-frequency spoken word (W), and $p(R_{\mathrm{HF}} \mid W_{\mathrm{LF}})$, the proportion of participants who "incorrectly" gave a high-frequency competitor response to the low-frequency (LF) spoken word. Likewise, $p(R_{\mathrm{LF}} \mid W_{\mathrm{LF}})$ corresponds to the proportion of participants who correctly gave a low-frequency target response to the low-frequency spoken word and $p(R_{\mathrm{LF}} \mid W_{\mathrm{HF}})$, the proportion of participants who incorrectly gave low-frequency competitor response to the high-frequency spoken word.

The rationale for subtracting frequency-matched target and competitor proportions was to yield an estimate of signal discriminability once frequency bias has been controlled for. At early gates, the sensory input should be characterized by a (near) complete ambiguity between target and competitor interpretations and choices are expected to be primarily driven by frequency biases; when these response biases are controlled for, the frequency-matched target-minus-competitor difference should be near zero. However, as gate duration increases, the target-competitor difference should increase, an indication that the input has started providing sensory evidence that the target, rather than the competitor, is being heard. Once disambiguation is complete, target-minus-competitor proportion differences should reach 1. Of interest here was the time interval, for each word pair, over which the target-competitor proportion difference changed from being larger than 0 to reaching 1. We will refer to this time interval as the *disambiguation window*, which we more precisely defined, for a given word pair, as the interval over which the target-competitor proportion difference became greater than 0.1 (and continued to increase thereafter) and when it first reached 0.9. The boundaries of the disambiguation windows were derived by linear interpolation on the function defined by the target-competitor proportion differences, as computed following Eq. (1), and gate durations.

The threshold value of 0.9, rather than 1 (indexing perfect performance), was adopted for practical reasons: Perfect target identification was not achieved even when

complete words were presented on 6 of the 28 pairs. On 3 of these items, errors consisted of misidentifying a high-frequency target as its low-frequency competitor. We attribute these errors to noise in response, due, at least in part, to participants mistyping their choice, rather than evidence that the signal continues to be ambiguous. Note that using a target-competitor proportion difference of 0.9 as the end of the disambiguation window ensures that for *all* stimuli, the spoken input provides reliable information available, which allows 29 listeners out of 30 to correctly identify the spoken word and reject its competitor. Furthermore, as discussed in the methods section, the use of a response proportion threshold based on both members of a pair ensures that disambiguating information has been reached for both the high-frequency and the low-frequency word even in cases where these words differ in terms of the speed at which such information becomes available.

Results revealed that, over the 28 word pairs, disambiguation windows extended for 103 ms on average (median = 105 ms), ranging from being quite brief (about 25 ms) to substantially long (about 215 ms). This analysis suggests that the recognition of a spoken word should not be conceived as a discrete moment, i.e., resulting from having access to a specific feature occurring at a well-defined point in time. Rather, we see a gradual shift, from a very small advantage for the target over the competitor, to a complete disambiguation, even when listeners are under no time pressure to respond and their perceptual choice at each gate probably reflects an asymptotic state. Fig. 2 illustrates the variability in disambiguation-window size by plotting the target-competitor proportion differences for three word pairs, one with a disambiguation window equivalent to the median size, among the 28 pairs (for the pair *slang–slak*, with a 110-ms window), one with a short disambiguation window (the pair *wolk–wolf*, with a 38.5-ms window), and one with a long disambiguation window (the pair *tand–tang*, with a 217-ms window). For the long-window pair, the target-competitor proportion difference steadily increased as gate duration increased, presumably due to gradually increasing nasalization differences in the vowels for *tand* and *tang*. By contrast, for the short-window pair, the difference remained close to 0 for an extended amount of time before increasing sharply and reaching asymptote very quickly. Here, the phonetic evidence discriminating *wolf* and *wolk* occurs as a relatively discrete event in time, with little prior evidence of the divergence. Overall, the spread of disambiguation window durations neatly illustrates the variability in the strength of coarticulation cues provided over time in normal speech.

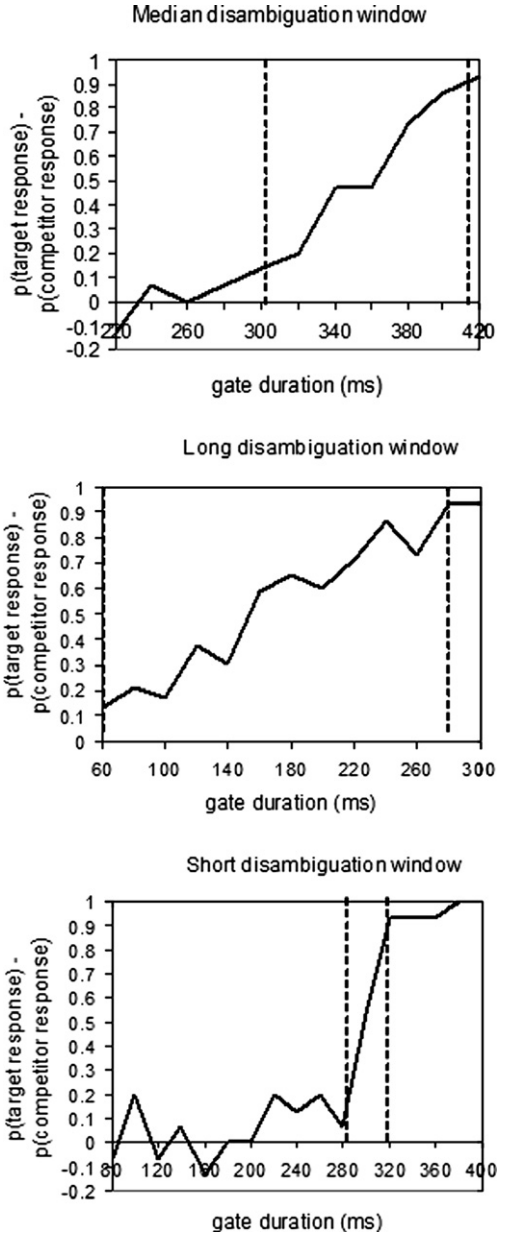

Fig. 2. Gating study. Difference between the proportion of target responses and the proportion of competitor responses as a function of gate duration for three word pairs which differ in terms of the duration of their disambiguation window (defined as the interval over which the target-competitor difference becomes larger than 0.1 and continues to increase until it reaches 0.9). Dashed lines indicate where the window boundaries are approximately located. See text for more details.

Eye-tracking

The data were first parsed into fixations and saccades. Saccade onsets and offsets were automatically detected using the thresholds for motion (0.2°), velocity (30°/s), and acceleration (8000°/s^2). Fixation duration corresponded to the time interval between successive saccades. Fixation location was assessed by averaging

the *x* and *y* coordinates of the fixation's samples, and by superimposing the fixation location onto the displayed grid and pictures. Fixations that fell within the grid cell containing a picture were hand-coded as fixations to that picture. All other fixations were coded as fixations to the grid, without further distinction. Fixations were coded from the beginning of the trial (i.e., the appearance of the display) until the target picture was fixated and clicked on.

Twenty experimental trials (accounting for 1% of the data) were excluded from the analyses because of poor calibration or track loss (5 trials), failure to fixate on the target object while or before clicking on it (10 trials) or selecting the wrong object (5 trials).

In order to visualize how fixations to target and competitor pictures changed over time, we computed, for each participant, the proportion of trials for which a given picture type or location (target, competitor, one of the two distractors, or the grid) was fixated, for each successive 10-ms time window from 0 to 1000 ms after target-word onset. When the last fixation of the trial, always to the target picture, ended before 1000 ms, its duration was extended. Proportions were then averaged over participants. Fig. 3 displays the fixation proportions to the target and competitor pictures in the high-frequency and low-frequency conditions. As a reference, the proportion of distractor fixations over time, averaged over both distractors and conditions, is also displayed. As apparent in the graph, the probability of fixating target and competitor pictures started to increase shortly after 250 ms after target-word onset, while the probability of fixating distractor pictures began to decrease. Furthermore, the probability of fixating high-frequency targets increased faster than the probability of fixating low-frequency targets. The probability of fixating competitor pictures revealed the inverse pattern, although the frequency effect was more modest. Nonetheless, the probability of fixating a high-frequency competitor picture remained higher than the probability of fixating a low-frequency competitor picture from about 600 ms until 900 ms after the onset of the spoken word. However, these fixation functions represent an average over word pairs that disambiguate at different moments in time. We return to the issue shortly.

Although fixation proportions give a good visual summary of the evolution of fixations to each picture type over time, they do not easily lend themselves to analyses. An important issue concerns the lack of independence between proportion values across time bins, given that one fixation, which generally lasts around 200 ms, contributes to several time bins simultaneously. Thus, treating time as an independent factor on proportions of fixations violates the assumption of independence. Solutions to this problem have been developed (see Tanenhaus, in press, for a review). However, another significant limitation of this data representation is that it does not capture well participants' actual behavior during a trial. A trial consists of a succession of fixations punctuated by saccades. We therefore chose to conduct analyses in terms of individual fixations, rather than proportions of fixations.

During the 500-ms preview, participants made 1.4 fixations on average (including the fixation that was tak-

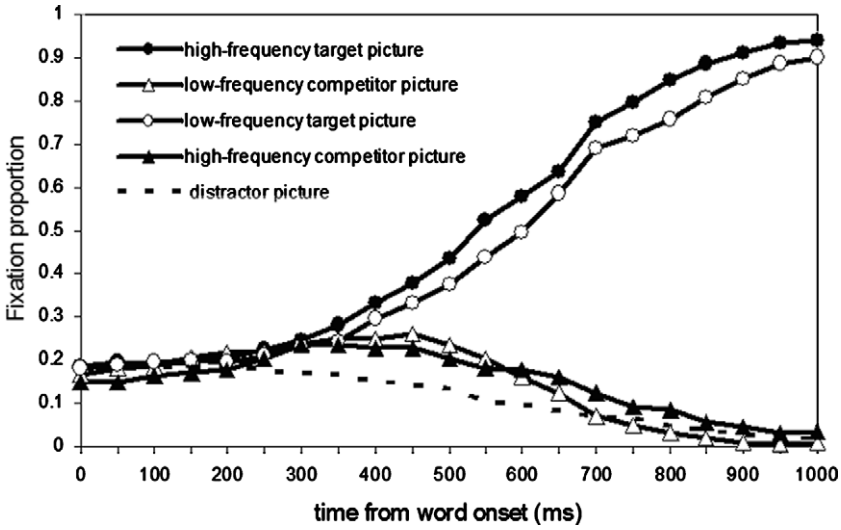

Fig. 3. Eye-tracking study. Proportion of fixations to high-frequency pictures (filled markers) low-frequency pictures (empty markers), and distractor picture (dashed line, no marker) over time. Fixations to target pictures are represented with circle markers, to competitor pictures, with triangle markers. To increase the clarity of the graph, the data were down-sampled and every 5 time bins are displayed.

ing place as the pictures appeared on the screen). The mean number of fixations taking place from the onset of the spoken word until the referent object was clicked with the mouse was 4.8, with more fixations in the low-frequency condition than in the high-frequency condition (5.2 vs. 4.5, $t(29) = 6.8$, $p < .001$, with a 95% confidence interval of ± 0.15). However, the number of distinct objects fixated did not differ (2.6 vs. 2.5, $t(29) = 1.3$, $p > .10$, with a 95% confidence interval of ± 0.08). Thus, participants do not appear to have made extra fixations randomly in the low-frequency condition; rather, people were more likely to return to an object previously fixated or make two or more successive fixations to the same object in the low-frequency condition (i.e., when the target was low-frequency and its competitor, high-frequency) than in the high-frequency condition. When specifically examining the fixations made to the competitor picture, the number of fixations per trial to a high-frequency competitor was greater than the number of fixations to a low-frequency competitor (0.61 vs. 0.51, $t(29) = 2.1$, $p < .05$, with a 95% confidence interval of $\pm .07$). By comparison, the number of fixations to the distractor picture (matched with the competitor picture for frequency) was lower and did not significantly vary as a function of its frequency (0.47 vs. 0.44, $t < 1$, with a 95% confidence interval of $\pm .07$).

Fixation analyses have demonstrated a frequency effect on competitor fixations. However, our focus here is to examine fixations to the competitor picture after a point at which the sensory input has disambiguated the competitor word from the target word. This point can be established, based on the gating data, as the end of the disambiguation window (i.e., at which the proportion of target responses minus that of competitor responses reached 0.9). This point, established for each word pair independently, was found to be 293 ms after the onset of the spoken word on average (ranging from 120 to 495 ms). Eye-movement data were realigned with respect to the end of the disambiguation window on a trial-by-trial basis. Fig. 4 displays the realigned data in terms of fixation proportions.

As apparent on the graph, the probability of fixating a high-frequency competitor picture declines more gradually than the probability of fixating a low-frequency competitor. Importantly here, the decline cannot be explained by averaging across word pairs that become disambiguated at various points in time because each trial was realigned to the point in time where virtually all gating responses identified the target picture as the referent to the spoken word.

Fixation analyses were conducted on the realigned data. Fig. 5 plots the number of fixations to the low- and high-frequency competitor that were initiated after the end of the disambiguating window, as a function of when, in fixation number, these fixations took place. For comparison, the numbers of fixations to the low- and high-frequency distractors are also displayed. Because the first fixations that were launched after the end of the disambiguating window (i.e., fixation 1) may have been planned based on sensory information before the end of the window, we focused on the fixations that were subsequently launched (i.e., fixations 2,

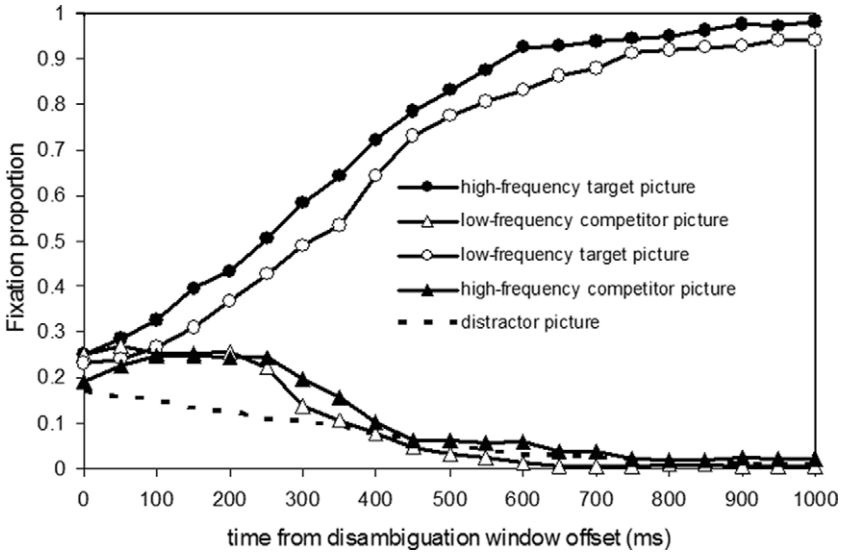

Fig. 4. Eye-tracking study. Proportion of fixations to high-frequency pictures (filled markers) low-frequency pictures (empty markers), and distractor picture (dashed line, no marker) over time, from the offset of the disambiguation window, as defined, for each word pair, from the gating data. Fixations to target pictures are represented with circle markers, and to competitor pictures, with triangle markers. To increase the clarity of the graph, the data were down-sampled and every 5 time bins are displayed.

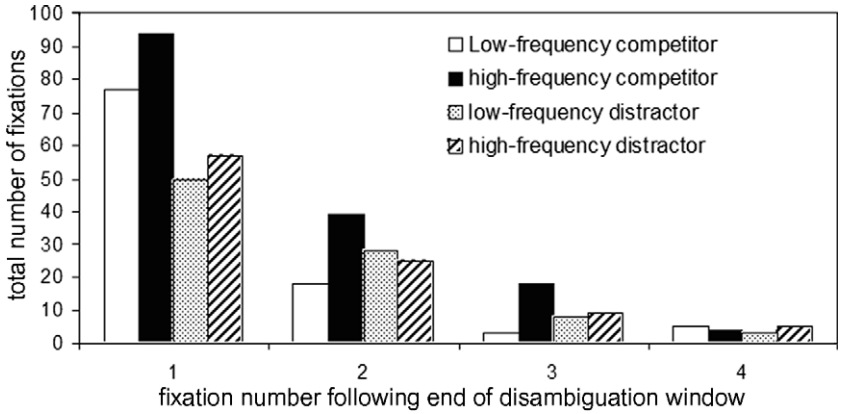

Fig. 5. Eye-tracking study. Number of fixations to the high-frequency and low-frequency competitor and distractor pictures as a function of the number of fixations following the end of the disambiguating window (see text for details).

3, and 4). Based on a binomial test, the number of fixations launched to the high-frequency competitor was significantly larger than that to the low-frequency competitor (fixation 2, $p < .005$; fixation 3, $p < .001$). No such difference was found on the fixations to the frequency-matched distractors.

This last set of analyses suggests that the high-frequency competitor continued to exert a stronger competition even after the signal had effectively disambiguated the target from the competitor, as indexed by gating responses. This result is consistent with a view in which signal processing is noisy and evolves dynamically, and in which information from the spoken input only gradually outweighs biases or priors. Because frequency-modulated fixations to competitors were observed after the end of the disambiguation window, as defined by gating responses, this late competition effect cannot be attributed to ambiguity in the signal, but to dynamical processing that has not reached a stable state when probed in real time.

Analysis of the link between lexical activation and visual search

Additional analyses were conducted to examine the process by which participants orient their gaze to spatial locations associated with pictured objects in the course of a trial, and how this process interacts with the development and evaluation of lexical hypotheses pertaining to the spoken input. In particular, we examined whether some spatial locations on the display tend to receive more fixations than others, and how this tendency interacts with the nature of the pictures that occupy these positions and evolves in the course of a trial. Dahan, Tanenhaus, and Salverda (in press) have shown that the first fixations of a trial tend to be directed to the pictures in the upper left and, to a lesser degree, the upper right, corners of the imaginary square. As the trial proceeds, the proportions of fixations to the other locations increase. The upper panel of Fig. 6 complements this finding by displaying the frequency with which each spatial location is visited for the first time in the course of a trial, as a function of the type of picture that occupies the location. As apparent on the graph, the number of "first" fixations differed across spatial locations ($\chi^2(3) = 25.5$ $p < .001$) and across picture types ($\chi^2(3) = 242.3$, $p < .001$). Furthermore, the proportion of "first" fixations to the target picture was smaller when the fixations were directed to the upper left location than when they were directed to the lower right location, with the other two locations showing intermediate proportions. This pattern of data was captured in a significant interaction between spatial location and picture type ($\chi^2(9) = 45.7$, $p < .001$). This suggests that fixations to the upper left location, which tended to occur early in the trial, were to a large extent unaffected by the identity of the picture that occupied the location; by contrast, the lower right location was mostly visited if the target picture occupied the location, or, we would argue, if the target picture did not occupy the other locations, visited earlier in the trial. The lower panel of Fig. 6 presents the frequency with which each spatial location was refixated within a trial. (A fixation was counted as a refixation only if other location(s) had been visited between the first fixation and its subsequent one). The rate of refixations was noticeably larger for some locations, i.e., those locations visited early in the trial, than others ($\chi^2(3) = 65.6$ $p < .001$), and target and competitor pictures were refixated more often than the distractors were ($\chi^2(3) = 535.5$ $p < .001$), with no significant interaction between these two factors ($\chi^2(9) = 6.7$, $p > .50$). This lack of interaction indicates that the proportion of refixations to the target and competitor pictures remains relatively stable across the spatial locations that these

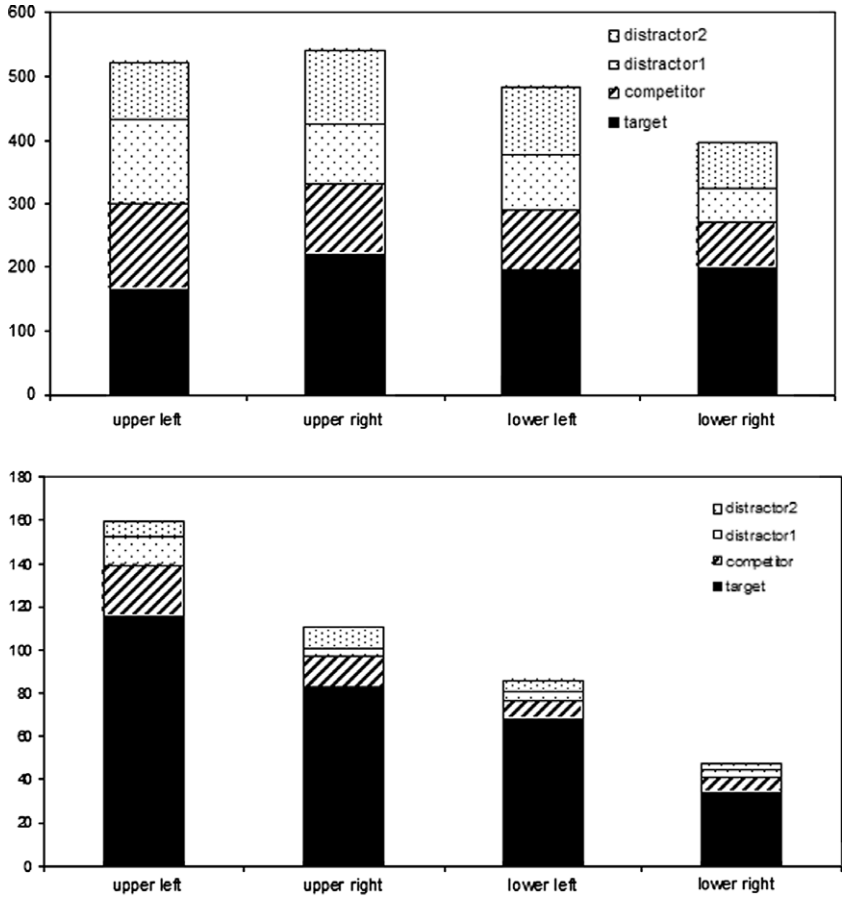

Fig. 6. Eye-tracking study. Number of "first" fixations (upper panel) and refixations (lower panel) as a function of the spatial location of the fixation and the picture type that occupied the location.

pictures may occupy once the likelihood that these locations have been visited earlier in the trial has been factored in.

Overall, these analyses indicate that participants' eye movements in this paradigm largely reflect the visual search people are engaged in as they accumulate visual and linguistic information about which of the four pictures is the referent. This finding suggests that the encoding of the pictures on the visual display may require some time before the representations resulting from this encoding can interact with the representations derived from the linguistic input. This finding also highlights the fact that eye movements in the visual-world paradigm do not solely reflect people's decision to orient their gaze toward potential referents in response to linguistic input, but also, to some degree, people's uptake of visual information.

In order to better describe how linguistic and visual information interact, we investigated whether having fixated or not a high-frequency or low-frequency competitor during preview would affect the likelihood of refixating it once the spoken input became available. (These analyses were originally reported in Dahan et al., in press.) Of the 840 trials (30 subjects × 28 experimental trials), 113 trials were excluded because the competitor picture was being fixated concurrently with the onset of the spoken input. For each of the remaining 727 trials, we determined whether or not the competitor had been fixated during preview and after spoken-word onset, separately for high-frequency and low-frequency competitors. Analyses on trials involving a high-frequency competitor revealed a significant interaction between whether or not the picture was fixated during preview and whether or not it was fixated after the spoken input began ($\chi^2_{(1)} = 5.2$, $p < .05$): Trials in which the high-frequency competitor had been fixated during preview and was refixated later in the trial were more frequent than the rate of fixating it during preview and after spoken-word onset would have predicted independently. Comparable analyses on fixations to the low-frequency competitors revealed no such interaction ($\chi^2_{(1)} = 1.3$, $p > .20$), with

a slight, non-significant, trend toward fewer trials with a refixation to the low-frequency competitor than without when the competitor had been fixated during preview. Thus, having fixated the high-frequency competitor picture during preview increased the likelihood of refixating it as the spoken input unfolded, whereas a fixation to the low-frequency competitor during preview had no significant impact on the likelihood of refixating it later in the trial.

This contingency analysis is important because it suggests that people do not simply redirect their attention toward any picture with a name that matches the spoken input (otherwise they would tend to refixate the low-frequency competitor as well). Rather, the results suggest that as the speech signal becomes available, some candidates become activated more than others (i.e., the high-frequency ones). If people happen to fixate the pictures associated with these high-frequency candidates, they will be more likely to refixate them than otherwise at a later point. By contrast, people do not tend to reorient their gaze toward the pictures associated with low-frequency candidates despite the fact that the names of these pictures are (at least temporarily) compatible with the spoken input. This finding suggests that the spoken input is evaluated with respect to lexical hypotheses, each one weighted by its frequency of occurrence in the language. This finding is incompatible with a view in which listeners bypass lexical processing and merely match the spoken input with the phonological string associated with each prefixated (and prenamed) object on the display.

General discussion

The combination of methods employed in this research provides valuable new data relevant to the understanding of the time course of ambiguity resolution in speech perception. The gating and eye-tracking tasks were selected because they both allow analysis of the perceptual impact of speech on a millisecond-by-millisecond basis, but they differ greatly in terms of the extent to which they reflect online perceptual processing. The gating technique provides absolute control over the duration of the speech signal that is presented to participants, but provides a working measure of the steady state of the word recognition system. This is because in gating listeners are encouraged to focus on accurate responses irrespective of the time taken to select a response. Other operational aspects of the task also ensured a measured response. Each trial involved the selection of both a target picture and a confidence rating, and this constant switching of tasks entailed a significant response-selection component, also extending the time available for processing of the signal. Thus, we can be confident that the majority of responses in gating reflected the state of the word recognition system when activation levels were approaching asymptote.

The gating data showed very clear effects of a frequency bias on responses at the early gates, where there was little or no information in the signal to discriminate between the two key words, replicating previous studies where no such restricted set of alternative responses was offered (Grosjean, 1980; Tyler, 1984). This advantage for high frequency candidates was eliminated by the time the complete spoken words were presented. More important for the current purposes are the analyses on individual item pairs. These analyses determined frequency-controlled disambiguation windows, which began at the point in time where the speech signal provided equal support for both lexical candidates and ended at a point where listeners were near unanimous in their selection of the correct target. Thus, the stretch of speech included in the disambiguation window for a particular item pair incorporated the critical information needed to discriminate between the two words.

On average, this stretch of speech lasted just over 100 ms, but in some cases was as little as 25 ms, and in others was as long as 215 ms. Thus, in some cases, speech can provide punctuate, discriminating events, much as the phonemic transcriptions of the words might suggest. Given a measured response, these short sections of speech can have strong effects on the state of the recognition system. More often, however, the evidence appears partial, and information must be accumulated over longer stretches in order to discriminate between lexical candidates with any degree of certainty. This fits in with numerous previous demonstrations of sensitivity to coarticulation in the perception of spoken words (e.g., Warren & Marslen-Wilson, 1987, 1988).

One possible objection to our estimates of the disambiguation windows is that the set of responses for each word pair comes from different individuals, who may be more or less conservative in the face of deterministic evidence from the signal. However, there was no indication that some listeners systematically identified the spoken word as the target earlier than others. In other words, the disambiguation windows did not appear to reflect responses averaged across fast and slow decision makers. Furthermore, variation in listeners' decision processes only would not easily account for the variability across word pairs. The data are most compatible with the notion of probabilistic information in the speech signal whose information value increases with increasing larger portion of speech.

We can also rule out the possibility that the variability in the temporal extent of the disambiguation window across item pairs reflects variability in the number and frequency of words that remained compatible with the spoken input after the competitor has been disfavored. Although these lexical aspects differed between item pairs, it is difficult to see how they could have any

impact on gating results given that participants gave almost exclusively target or competitor responses. Consequently, a decrease in the rate of one type of responses brought a proportionally equivalent increase in the rate of the other type, with no apparent influence of other, non-depicted alternatives.

Another possible objection to our long disambiguation window estimates hinges on the fact that each window was established based on responses made to two words, the high-frequency and low-frequency items of each pair. Recall that the temporal extent of the disambiguation windows was based on the computation of target-competitor proportion differences; the target proportion corresponded to the proportion of participants who correctly identified the spoken word they heard; the competitor proportion corresponded to the proportion of participants who gave the same (but here incorrect) response when hearing the other item of the pair. Some disambiguation windows may appear to extend for a substantial amount of time in spite of the presence of punctuate and deterministic phonetic information in each word if the information providing support to one interpretation over the other became available early for one word but late for the other. Thus, the proportion of target responses would start to increase early because of one of the words of a pair but would not reach 0.9 before a substantial delay because of the late recognition of the other word of the pair. Although this possibility is difficult to rule out, it is unlikely to have extended the estimate of some disambiguation windows by a large amount because the two items of each pair were closely matched in terms of the properties of their overlapping information. Thus, we argue, most of the variability across item pairs reflects differences in how information must accumulate in order to discriminate between lexical candidates with any degree of certainty.

It is possible that the gating procedure adopted here, in which increasingly longer portions of a spoken word were presented and listeners' responses selected from a closed set of four alternatives, may have led people to persist in their initial choice and require more evidence before altering their decision than they would have otherwise. Thus, the temporal extent of the disambiguation windows (but not the variability across item pairs) may be over-estimated. Our rationale for adopting this procedure was that it matched the procedure of the eye-tracking study most closely. Furthermore, any factor that leads us to overestimate the disambiguation point can only make our key test—of whether online frequency effects extend beyond disambiguation point—more conservative.

Whereas gating tells us primarily about the unfolding of the signal, eye movements in the visual-world paradigm are influenced by several factors. Alongside the unfolding of the signal, listeners' fixations to visual targets are influenced by the dynamics of the word-recognition process and by the constraints of the visual system, as well as the interconnection between the two. These factors combined produced a greatly extended period of time in which word frequency affected the probability of fixating a competitor item. However, the relative contributions of the signal, the perceptual system and the visual system were not clear from this means of analysis. The advantage of collecting gating data on the same stimulus set is that the influence of the different factors contributing to visual-world fixation patterns could then be isolated.

This was the rationale behind the analysis underlying Fig. 4. Here, the origin of the x-axis is an estimate, gleaned from the gating data, of the time at which the signal becomes unambiguous. Further analysis of these realigned data minimized the influence of any lag due to programming of a saccade by only considering the second and third fixations initiated after the offset of the disambiguation window (see Fig. 5). This analysis revealed that, after this point, listeners oriented their gaze toward the high-frequency competitor more often than toward the low-frequency competitor or the distractor. So the key contribution of the research is to demonstrate that the perceptual process in speech is a significant limiting factor on recognition speed, over and above the ambiguity of the signal. This result is in accord with previous hints from semantic priming studies using word fragments (Gaskell & Marslen-Wilson, 2002; Zwitserlood & Schriefers, 1995), and eye-tracking experiments (Dahan et al., 2001). However, none of the previous studies have so successfully teased apart the relative contributions of signal ambiguity and processing dynamics.

Our results are at odds with previous research suggesting that frequency effects are limited to regions of signal ambiguity in the speech waveform (e.g., Marslen-Wilson, 1990) but fit better with the body of literature discussed in the introduction (cf. Luce & Pisoni, 1998) demonstrating frequency effects of competitors on the recognition of complete spoken words. What should we make of evidence that competition can extend over a substantial period, even after the signal has provided disambiguating information? Might speech perception be less optimal than Marslen-Wilson (1984) claimed it to be? This is one potential conclusion that could be drawn from the current research. Our data could be interpreted as suggesting that frequency biases are unduly influential even in the face of clear phonetic evidence. In other words, even when there is strong bottom-up evidence in favor of a low frequency word, there remains an enduring tendency to entertain an alternative hypothesis even though it now has a worse fit with the speech information. This over-commitment to an earlier strong candidate could

be explained in terms of an inability to reevaluate previous hypotheses in the face of new sensory evidence. However, there is an alternative means of explaining the data that preserves the optimality principle. Following the recently developed framework of rational analysis (Anderson, 1990; Oaksford & Chater, 1998; see Chater & Oaksford, 1999; for a brief overview), optimality can be evaluated with respect to the goals of the system, the environment in which it operates, and, critically, its computational limitations. If we assume either that there is noise in the perceptual system, or that the sampling of the speech input is capacity-limited, then an optimal system in Bayesian terms should integrate the signal with prior biases in just the way we see here: gradually, and over an extended period (cf. Norris, 2006). Further research is essential if we are to go beyond merely classifying the spoken-word recognition process as qualitatively consistent with Bayesian optimality. To this end, recent progress has been made in other areas of perception in terms of confirming quantitative predictions of the Bayesian approach as applied to cue integration (e.g., Ernst & Banks, 2002; Jacobs, 2002; Tassinari, Hudson, & Landy, 2006). Thus it seems plausible that, as in other areas, the spoken-word recognition system is responding optimally given certain external and/or internal constraints.

We should clarify that the extended influence of competitor frequency that we have observed does not suggest that the perceptual system is slow to respond to speech in its entirety. Many studies, particularly those involving eye-tracking and event-related potential methods (e.g., Dahan et al., 2001; Dahan & Tanenhaus, 2004; Sanders & Neville, 2003; Van Petten, Coulson, Rubin, Plante, & Parks, 1999), have demonstrated that the perceptual system responds swiftly to changes in the speech signal, and our own data bear this out. Nonetheless, the swift initial response of the system is only part of the story, and the focus of our paper is on the tail of this response, showing that changes in the speech signal have near-immediate perceptual consequences, but that the new information is only fully integrated with prior information over a considerable length of time. Recent data from a very different methodology (the psychological refractory period) point to an equivalent extended integration process for perception of phonemic information (Gaskell, Quinlan, Tamminen and Cleland, submitted).

A further aim of our study, especially relevant to the theme of this special issue, was to examine an aspect of the visual-world paradigm that has not yet been properly fleshed out. Participants' eye movements largely reflect the visual search people are engaged in as they accumulate visual and linguistic information about which of the four pictures is the referent. The precise nature and timing of the interaction between these two types of information is likely to affect how lexical activation translates into eye movements, but a full working model of the interplay between these factors has not yet been elaborated. In the present study, initial fixations appeared to be determined by spatial locations more than by the picture type that occupied these locations. As the trial progressed and people gathered more information about the display, the influence of linguistic input on eye movements became stronger, and subtle effects, such as the competitor frequency on the probability of refixating on a picture, emerged.

The implications of these analyses are significant as they may help explain why the frequency effect reported here appeared to take place at a later time, with respect to the onset of the spoken word, than the frequency effect previously reported with the visual-world paradigm (Dahan et al., 2001). In one of the experiments reported there, a low-frequency target was presented along with two onset-overlapping competitors, one of high frequency and the other of low frequency. Probabilities of fixation over time revealed an early bias toward the high-frequency competitor, most visible between roughly 200 and 450 ms after spoken-word onset (see Dahan et al., 2001; Fig. 2). In the current study, the effect of frequency, as graphically depicted in the difference between low-frequency and high-frequency targets and competitors in Fig. 2, expressed itself later in time.

At least one difference between the two studies may explain this contrast. In the Dahan et al. (2001) study, participants benefited from about 900 ms of preview before the onset of the referent's name (i.e., 500 ms of silence followed by the carrier instruction "Pick up the...", which was on average 402 ms long). Longer preview provides potentially more time for participants to inspect the display and gain knowledge about the conceptual characteristics associated with the four critical spatial locations. As our contingency analyses have shown, this knowledge has an impact on subsequent fixations, with a greater chance of refixating the high-frequency competitor than the overall rate of fixating the picture would predict. This suggests that the more opportunities to fixate on the high-frequency competitor participants have had before hearing the name of the referent picture, the more likely they are to direct their gaze toward that picture once spoken input becomes available. Under these conditions, an early effect of lexical frequency can be observed; when only limited opportunities to inspect the display before the spoken input begins are provided, frequency appears to affect fixations substantially later in the trial. This pattern may lead to the erroneous conclusion that frequency affects lexical activation after some delay.

We cannot pretend, at this stage, that there exists a precise specification of how lexical processing links into and informs visual search in the course of a trial. The argument above merely highlights the increasing need to provide such a model. Eye-tracking has become a valuable means of adding to our understanding of speech perception and spoken-word recognition, complementing other existing methodologies. Nonetheless, the visual-world paradigm will only achieve its full potential if we are able to gain a better understanding of how language perception and visual processes combine in real time.

Conclusions

The major contribution of our study is to demonstrate that the dynamics of the perceptual process itself is a significant limiting factor on recognition speed, over and above the ambiguity of the signal. While the notion of processing-internal dynamics has been widely accepted in related fields, the spoken-word recognition community may have focused too much on the external sequentiality of the spoken input, to the detriment of its internal sequentiality. Our results highlight the influence of both of these factors, and bring the internal dynamics of the perceptual process itself to the attention of the growing number researchers who, assisted in part by the development of new experimental techniques (e.g., eye-tracking, event-related brain potentials), aim to understand how speech is processed and interpreted in realtime.

Acknowledgments

This work was conducted at the Max Planck Institute for Psycholinguistics (Nijmegen, The Netherlands). We wish to acknowledge the support of the Max Planck Society, of the National Science Foundation under Grant No. 0433567, the National Institutes of Health (1 R01 HD 049742-1) and the UK Medical Research Council (G0000071). We also thank Ted Strauss and Jim Magnuson for their help in conducting simulations with jTRACE.

Appendix A

High-frequency items			Low-frequency items			Number of onset-overlapping phonemes
Word with English translation and log frequency			Word with English translation and log frequency			
bed	bed	2.5	bel	bell	1.5	2
bezem	broom	0.6	bever	beaver	0.0	2
boek	book	2.6	boei	buoy	0.6	2
computer	computer	1.7	komkommer	cucumber	0.6	3
fles	bottle	2.0	flat	apartment building	1.5	3
haak	hook	1.3	haai	shark	0.5	2
hand	hand	3.0	ham	ham	1.2	2
hart	heart	2.3	harp	harp	0.3	3
hoed	hat	1.6	hoef	hoof	0.3	2
kado	present	1.3	kanon	cannon	1.0	2
kandelaar	candelabra	0.5	kangoeroe	kangaroo	0.0	2
kat	cat	1.9	kam	pitcher	0.9	2
kerk	church	2.3	kers	cherry	0.7	3
koffie	coffee	2.0	koffer	suitcase	1.7	3
krant	newspaper	2.1	krans	wreath	0.8	4
kroon	crown	1.4	kraan	faucet	1.1	2
plant	plant	1.9	plank	wood board	1.5	4
pot	jar	1.5	pop	doll	1.3	2
schaap	sheep	1.4	schaats	ice skate	0.0	3
schip	ship	2.1	schild	shield	0.8	3
slang	snake	1.4	slak	snail	0.7	3
slot	lock	1.9	slee	sled	0.3	2
ster	star	1.8	step	scooter	0.0	3
tand	tooth	1.9	tang	pliers	0.7	2
ton	barrel	1.5	tol	top	0.7	2
vork	fork	1.1	vos	fox	0.8	2
wiel	wheel	1.3	wieg	cradle	1.0	2
wolk	cloud	1.7	wolf	wolf	1.2	3

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.

Anderson, J. R. (1990). *The adaptive character of thoughts*. Hillsdale, NJ: Erlbaum.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). The CELEX lexical database (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3*, 57–65.

Cooper, R. (1974). The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84–107.

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). Continuous mapping from sound to meaning in spoken-language comprehension: immediate effects of verb-based thematic constraints. *Language and Cognitive Processes, 30*, 498–513.

Dahan, D., Tanenhaus, M. K., & Salverda, A.P. (in press). The influence of visual processing on phonetically driven saccades in the ''visual world'' paradigm. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), Eye-movements: A window on mind and brain. Oxford, England: Elsevier.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429–433.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology, 45*, 220–266.

Gaskell, M. G., Quinlan, P.T., Tamminen, J.T., & Cleland, A.A. (submitted for publication). The nature of phoneme representation in spoken word recognition.

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

Hallett, P. E. (1986). Eye movements. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance* (pp. 10-1–10-112). New-York: Wiley.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences, 6*, 345–350.

Jurafsky, D., Bell, A., & Girand, C. (2002). The role of the lemma in form variation. In C. Gussenhoven & N. Warner (Eds.). *Laboratory Phonology* (7, pp. 3–34). Berlin: Mouton de Gruyter.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: a parallel distributed processing account. *Journal of Memory and Language, 32*, 474–516.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. Research on Speech Perception, Technical Report No. 6. Indiana University. Bloomington, IN.

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes, 16*, 565–581.

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

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244*, 522–523.

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science, 189*, 226–227.

Marslen-Wilson, W. (1984). Function and process in spoken word recognition. In H. Bouma & D. G. Bouwhis (Eds.), *Attention and performance X: Control of language processes* (pp. 125–150). Hillsdale, NJ: Lawrence Erlbaum Associates.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71–102.

Marslen-Wilson, W. D. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 148–172). Cambridge, MA: MIT Press.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representations and process in lexical access: words, phonemes, features. *Psychological Review, 101*, 653–675.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29–63.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: The MIT Press.

McClelland, J. L. (1993). Toward a theory of information processing in graded, random, and interactive networks. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 655–688). Cambridge, MA: MIT Press.

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

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception & Performance, 25*, 1363–1389.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition, 52*, 189–234.

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327–357.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behavioral & Brain Sciences, 23*, 299–370.

Oaksford, M., & Chater, N. (1998). *Rational models of cognition*. Oxford, England: Oxford University Press.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–109.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*, 261–300.

Sanders, L. D., & Neville, H. J. (2003). An ERP study of continuous speech processing I. Segmentation, semantics, and syntax in native speakers. *Cognitive Brain Research, 15*, 228–240.

Strauss, T. J., Harris, D., & Magnuson, J. S. (in press). jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. Behavior Research Methods, Instruments, and Computers.

Tanenhaus, M. K. (in press). Eye movements and spoken language processing. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), Eye-movements: A window on mind and brain. Oxford, England: Elsevier.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science, 268*, 1632–1634.

Tassinari, H., Hudson, T. E., & Landy, M. S. (2006). Combining priors and noisy visual cues in a rapid pointing task. *Journal of Neuroscience, 26*, 10154–10163.

Tyler, L. K. (1984). The structure of the initial cohort: evidence from gating. *Perception and Psychophysics, 36*, 417–427.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review, 108*, 550–592.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 394–417.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: levels of processing in perception of spoken words. *Psychological Science, 9*, 325–329.

Wagenmakers, E. J., Steyvers, M., Raaijmakers, J. G., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology, 48*, 332–367.

Warren, P., & Marslen-Wilson, W. D. (1987). Continuous uptake of acoustic cues in spoken word-recognition. *Perception and Psychophysics, 41*, 262–275.

Warren, P., & Marslen-Wilson, W. D. (1988). Cues to lexical choice: discriminating place and voice. *Perception and Psychophysics, 43*, 21–30.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics, 35*, 49–64.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32*, 25–64.

Zwitserlood, P., & Schriefers, H. (1995). Effects of sensory information and processing time in spoken-word recognition. *Language and Cognitive Processes, 10*, 121–136.