Perception & Psychophysics
2006, 68 (1), 1-16

Segment duration as a cue to word boundaries in spoken-word recognition

KEREN B. SHATZMAN and JAMES M. McQUEEN
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

In two eye-tracking experiments, we examined the degree to which listeners use acoustic cues to word boundaries. Dutch participants listened to ambiguous sentences in which stop-initial words (e.g., *pot*, jar) were preceded by *eens* (once); the sentences could thus also refer to cluster-initial words (e.g., *een spot*, a spotlight). The participants made fewer fixations to target pictures (e.g., a jar) when the target and the preceding [s] were replaced by a recording of the cluster-initial word than when they were spliced from another token of the target-bearing sentence (Experiment 1). Although acoustic analyses revealed several differences between the two recordings, only [s] duration correlated with the participants' fixations (more target fixations for shorter [s]s). Thus, we found that listeners apparently do not use all available acoustic differences equally. In Experiment 2, the participants made more fixations to target pictures when the [s] was shortened than when it was lengthened. Utterance interpretation can therefore be influenced by individual segment duration alone.

"Robin and Chris had once paid for all the gardening work." If this sentence were spoken, uncertainty could arise as to whether the speaker said "once paid" or "one spade." This is because, unlike printed language, in which the beginnings and ends of words are unambiguously marked with blank spaces, spoken language does not typically have clear breaks between words. In the absence of clear word boundaries in the speech signal, lexical ambiguities can arise. Of course, completely ambiguous sentences such as these are not common. However, ambiguity resolution is required for any given spoken sentence. As speech unfolds over time, words that are fully or partially consistent with the input become activated and compete with one another. A certain degree of ambiguity is therefore present, at least temporarily, in all sentences. However, the competition process resolves these ambiguities, such that the result of the recognition process is a parse of nonoverlapping words. This activation and competition process is instantiated in several current models of spoken-word recognition, including TRACE (McClelland & Elman, 1986), Shortlist (Norris, 1994), and the distributed cohort model (DCM; Gaskell & Marslen-Wilson, 1997).

The information in the acoustic signal is, of course, the most important influence on the lexical competition process. One source of acoustic information is that which marks word boundaries. Explicit physical cues for word onset include glottal stops and laryngealized voicing for vowel-initial words and increased aspiration on voiceless stops (Christie, 1974; Lehiste, 1960; Nakatani & Dukes, 1977). Note that the usage of such cues is perfectly compatible with competition-based recognition: Cues provide likely locations for word boundaries, thereby modulating the competition process (Norris, McQueen, Cutler, & Butterfield, 1997). The general mechanism of lexical competition is necessary, because, while explicit cues may mark some word boundaries in the speech signal, these cues are not always present.

Word onsets can also be marked by prosodic cues, such as duration, amplitude, and pitch. Acoustic-phonetic research has revealed differences in articulatory and acoustic properties of segments and syllables, depending on the location of word boundaries. For instance, Turk and Shattuck-Hufnagel (2000) compared the durations of syllables in triads such as *tune acquire*, *tuna choir*, and *tune a choir*. They found that the location of word boundaries influenced the duration patterns of the syllables. For example, the sequence /tjuːn/ was found to be longer in *tune acquire* than in *tuna choir*. Segment duration also depends on the position of the segment with respect to word boundaries. Segments in word-initial position, for example, tend to be longer than those in word-medial or word-final position (see, e.g., Klatt, 1974; Oller, 1973; Umeda, 1977).

These systematic variations in the productions of segments have been incorporated into a general framework using the notion of the prosodic hierarchy—the view that spoken utterances are hierarchically organized, with large prosodic constituents, or domains, consisting of smaller constituents (see, e.g., Beckman & Pierrehumbert, 1986; Nespor & Vogel, 1986; see Shattuck-Hufnagel & Turk, 1996, for a review). Recent studies (Cho & Keating, 2001; Fougeron, 2001; Fougeron & Keating, 1997) have shown that segments in initial position in higher level constituents are different, articulatorily and acoustically, from

The authors thank Delphine Dahan for helpful discussions, Loulou Edelman for lending her voice for the recordings, and Steve Goldinger and two anonymous reviewers for their comments on previous versions of the manuscript. Correspondence should be addressed to K. B. Shatzman, who is now at Utrecht Institute of Linguistics OTS, Trans 10, 3512 JK Utrecht, The Netherlands (e-mail: keren.shatzman@let.uu.nl).

Copyright 2006 Psychonomic Society, Inc.

initial segments in lower domains. Ipso facto, within a prosodic domain, a domain-initial segment has different fine-grained phonetic properties from a domain-medial segment. However, there is great variability among speakers in how many and which domains they distinguish (Fougeron & Keating, 1997). Furthermore, speakers vary with regard to the precise phonetic differences they produce to distinguish between contrasting boundary positions (Barry, 1981; Quené, 1992). That is, different speakers may exhibit different boundary phenomena. Due to this variation, fine-grained phonetic properties, on their own, seem insufficient to solve the segmentation problem.

Nevertheless, perceptual studies have shown that listeners can use these fine-grained acoustic differences, when they are present, to help in finding word boundaries. Davis, Marslen-Wilson, and Gaskell (2002) investigated the temporary ambiguity that arises due to initially embedded words (such as *cap* in *captain*). Using a cross-modal identity-priming task, Davis et al. compared the activation of both the shorter and the longer words in sentences in which the speaker intended the longer word (e.g., *captain*) and in sentences in which the speaker intended the shorter word (e.g., *cap*). Their results showed that there was more activation of the shorter word (*cap*) when the ambiguous sequence /kæp/ came from a monosyllabic word than when it came from the longer word (*captain*), and there was more activation for the longer word when the sequence came from a longer word than when it came from a shorter word. Acoustic analyses of the stimuli indicated that the ambiguous sequence was longer when it was a monosyllabic word than when it corresponded to the initial syllable of the longer word. Using eye movement data, Salverda, Dahan, and McQueen (2003) demonstrated that the duration of the ambiguous sequence in the case of initially embedded words in Dutch (e.g., *ham* in *hamster*) can modulate the amount of transitory fixations to pictures representing the monosyllabic embedded words. By manipulating the duration of the initial syllable of the longer words, they showed that longer sequences generated more monosyllabic word interpretations, indicating that listeners use fine-grained information to bias their lexical interpretations of utterances.

Recently, in a study in French (Christophe, Peperkamp, Pallier, Block, & Mehler, 2004), listeners were presented with sentences containing a local lexical ambiguity. For example, the phrase *chat grincheux* (grumpy cat) contains the word *chagrin* (sorrow). Listeners were delayed in recognizing the word *chat* in these sentences, relative to sentences in which there was no local lexical ambiguity. However, there was no such delay when there was a phonological phrase boundary between the two words containing the local lexical ambiguity. This again demonstrates that listeners exploit the prosodic structure of utterances in the online segmentation of continuous speech.

The influence of individual segment duration on listeners' offline segmentation judgments has been demonstrated in a study by Quené (1992). Using ambiguous two-word sequences such as the Dutch phrases *diep in* (deep in) and *die pin* (that pin) in a forced choice experiment, he showed that Dutch listeners made use of the duration of the intervocalic consonant in segmenting these word pairs. The study showed that manipulating the duration of this intervocalic consonant influenced listeners' explicit lexical segmentation judgments. Similarly, in a recent study in Dutch (Kemps, 2004), participants were exposed to an ambiguous sequence in which the consonant [s], appearing as the onset of a word, could also function as the plural suffix of the previous word (e.g., *kerel soms* [guy sometimes] could also be *kerels soms* [guys sometimes]). Participants' forced choice judgments in a number decision task showed that they were attending to the duration of the [s] to resolve the ambiguity between the two possible interpretations.

Studies using online measures have more directly examined the effect of segment duration on word recognition in continuous speech. Gow (2002), using the cross-modal priming paradigm, looked at phrases that were ambiguous due to the phonological process of place assimilation, such as the phrase *right berries*, which could also be produced sounding like *ripe berries*. Participants appeared to be able to discriminate modified and unmodified forms on the basis of acoustic information (i.e., subtle differences between the assimilated word-final stop in *right* [raɪp] *berries* and a genuine [p] in *ripe berries*), even for tokens that were judged to be highly ambiguous in an offline perceptual rating task.

Gow and Gordon (1995) examined recognition of lexically ambiguous sequences that could be interpreted as either two words or one longer word (e.g., *two lips*/*tulips*), again using cross-modal priming. They found priming of responses to the second word (e.g., *lips*) when it had just been heard as part of the two-word sequences (*two lips*) but not when it was part of the single-word sequences (*tulips*). The word-initial consonants (e.g., the [l] in *two lips*) were longer than the noninitial consonants (e.g., the [l] in *tulips*). Gow and Gordon concluded that listeners were using this durational cue in lexical access and segmentation. A similar priming study by Spinelli, McQueen, and Cutler (2003) examined segmentation of lexically ambiguous sequences in French. Specifically, they investigated the case of liaison, a process in which, during resyllabification across word boundaries, an otherwise silent consonant is realized by the speaker. In *dernier oignon* (last onion), for example, the final [ʁ] of *dernier* is produced and resyllabified with the following syllable, making the phrase sound like *dernier rognon* (last kidney). French listeners' segmentation of such ambiguous liaison phrases appeared to be influenced by the duration of the critical consonant: The word-initial consonants were longer than those in the liaison environments (e.g., [ʁ] in *dernier rognon* vs. *dernier oignon*). But neither Gow and Gordon nor Spinelli et al. demonstrated that the duration of the critical consonants was the factor that actually guided the listeners' segmentation. That is, it was not shown that manipulation of the critical consonant's duration alone influenced segmentation. In addition, other cues to word boundaries were not examined. The influence of other acoustic correlates of word boundaries therefore remains uncertain in these studies. Consequently, attributing the perceptual effect found in these studies to the acoustic cue of word-initial segment duration, though plausible, is somewhat conjectural.

The goal of the present study was to examine the degree to which different acoustic cues to word boundaries are used by listeners in their online segmentation of continuous speech. Previous studies involving lexically ambiguous phrases have tended to use segmentally heterogeneous item sets, making it impossible to draw strong inferences about exactly which acoustic properties of the speech materials were determining listener behavior. For example, in the Spinelli et al. (2003) study, the liaison consonant was [ʁ], [p], [t], [n], or [g]. Due to this kind of diversity, it would be impossible to conduct one and the same detailed acoustic analysis across all of the materials in such studies. Consequently, it is also not possible to examine, in a single analysis of the full set of materials, the extent to which the acoustic measurements correlate with the perceptual effect. In the present study, therefore, we used such phrases as "one spade" and "once paid," in which ambiguity occurs regarding whether the phrase contains a cluster-initial word or a word-final [s] followed by a word beginning with a singleton consonant. Thus, the segmental content of the ambiguous phrases was controlled, enabling both a detailed acoustic analysis and a direct test of whether particular aspects of acoustic-phonetic detail influence listener performance. Because of the homogeneity of our item set, all items were subject to the same acoustic analysis, and the measurements of this analysis could be correlated with listeners' behavior to determine the extent to which each acoustic cue might have influenced that behavior.

A Dutch speaker produced Dutch sentences that contained a stop-initial word (e.g., the word *pot* in *ze heeft wel eens pot gezegd* [she said once jar]) or matched sentences that contained a cluster-initial word that consisted of the stop-initial word and the preceding [s] (e.g., the word *spot* in *ze heeft wel een spot gezegd* [she did say a spotlight]). Thus, the sentences differed in their precise acoustic-phonetic realization but were phonemically identical. The degree to which a stop-initial word in this context can be discriminated from a cluster-initial word should depend on the acoustic correlates of word boundaries. Acoustic measurements of the ambiguous sequences (e.g., *eens pot* vs. *een spot*) were performed to assess the differences between them.

Differences between the acoustic properties of the ambiguous sequences are effective cues, however, only to the extent that listeners can perceive these differences and use them in online segmentation and word activation. Note that, although studies such as those of Quené (1992) and Kemps (2004) suggest that listeners are sensitive to fine-grained durational differences in the speech signal, because these studies used offline measures (i.e., forced choice tasks), they do not show that listeners use such differences during normal speech processing.

In the present research, we therefore used the eye-tracking paradigm to evaluate listeners' ability to distinguish between the two readings of the ambiguous sentences. This paradigm has been used to study the time course of lexical access (Allopenna, Magnuson, & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; for an overview of the paradigm, see Tanenhaus & Spivey-Knowlton, 1996). Several researchers (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Salverda et al., 2003) have also demonstrated that this paradigm can be used to examine the modulation of lexical activation of potential candidates over time at a very fine-grained level. In the standard form of the eye-tracking paradigm, participants are shown four pictures on a computer screen as they hear a spoken sentence. They are instructed to use the computer mouse to click on and move the picture of the object referred to in the sentence. The probability of fixating a pictured object has been shown to vary with the goodness of fit between the name of the picture and the spoken input.

In the present study, we manipulated the spoken input by cross-splicing in order to evaluate the effect of the realization of the ambiguous sequence on lexical activation as reflected by the participants' fixations to the pictured objects. In Experiment 1, the target word and the preceding segment [s] (e.g., the [s] and the word *pot* in *ze heeft wel eens pot gezegd*) were replaced either by the cluster-initial word (*spot*) or by the target word (*pot*) and the preceding [s] from another recording of the sentence. Of primary interest was whether the participants' fixations to the target picture differed across the splicing conditions. Subsequently, we examined which acoustic information the participants might be using by correlating their performance in the eye-tracking task with the differences found in the acoustic analyses.

EXPERIMENT 1

Method

Participants. Twenty-four student volunteers from the Max Planck Institute subject pool took part in this experiment. They were all native speakers of Dutch. They were paid for their participation.

Materials. The target words consisted of 20 stop-initial Dutch nouns referring to picturable objects (e.g., *pot*). The target words were chosen such that the addition of an initial [s] to each word would result in another Dutch noun. For example, adding an initial [s] to the Dutch word *pot* makes the word *spot*. Note that the cluster-initial counterpart words were not necessarily picturable nouns. We intentionally avoided a design in which the target's cluster-initial counterpart would be present on the screen, because having two possible referents would be likely to elicit confusion and induce the participants to adopt a strategy to deal with those items. Each target was instead paired with a picturable noun that had the same initial two-consonant cluster as the target's cluster-initial counterpart. For instance, the target *pot* was paired with *spin* [spider] (the initial cluster of *spin* was thus matched to that of *spot*). We will refer to the cluster-initial picturable noun as the *competitor*. There were no semantic or morphological relationships between the target and competitor words within each pair. Two additional picturable nouns were assigned to each target and competitor pair (e.g., *vuur* [fire] and *kompas* [compass]). These distractors were semantically and phonologically unrelated to the target, the competitor, or the target's cluster-initial counterpart. The full set of items is presented in the

Appendix. Line drawings of the items were selected from a number of picture databases (including the picture sets from Cycowicz, Friedman, Rothstein, & Snodgrass, 1997, and Snodgrass & Vanderwart, 1980, as well as the Art Explosion Library, 1995).[1]

Two recording contexts were constructed for each experimental item. One of the contexts referred to the target word, and the other referred to the target's cluster-initial counterpart. The sequences containing the target or its counterpart were identical and, therefore, fully ambiguous (e.g., *ze heeft wel eens pot gezegd* is phonemically identical to *ze heeft wel een spot gezegd*). The two words preceding the target and its counterpart were always *wel een(s)*.

A female speaker of Dutch who was naive to the purpose of the experiment read the sentences aloud in a sound-attenuated booth in random order. Recordings were made on a DAT tape (sampling at 48 kHz with 16-bit resolution). All sentences were recorded a minimum of four times. They were then redigitized at a sample rate of 16 kHz and manipulated using speech-editing software (Xwaves). For each target word, two spliced sentences were created. The carrier phrase for both versions consisted of the initial portion of the target-bearing sentence (up to but not including the [s]; e.g., *ze heeft wel een*) and the final portion of the same sentence (e.g., *gezegd*). In one version (hereafter, the *identity-spliced version*), the target word (e.g., *pot*) and the preceding [s] were taken from another token of the target recording context and spliced into the carrier phrase. In the other version (hereafter, the *cross-spliced version*), the target and the preceding [s] originated from the cluster-initial recording context (e.g., *spot*; see Table 1). The cross- and identity-spliced sentences were thus lexically identical but differed in the origin of the ambiguous sequence (i.e., whether this sequence was taken from the target or the cluster-initial recording context). All splicing points were at zero crossings, and care was taken to avoid any acoustic artifacts, such as clicks or other distortions.

In addition to the 20 experimental items, 50 sets of fillers were constructed. For each filler trial, a picturable word was selected to play the role of the target, along with three picturable distractor words. Pictures for the filler trials were selected from the same databases as were used for the experimental trials. The aim of the filler trials was to prevent the participants from developing expectations regarding the likelihood of a picture to be the target. Specifically, in all experimental trials, the target word started with either a [p] or a [t]. Additionally, the initial portion of the carrier phrase in these trials was always very similar (e.g., *ze heeft wel een*). Thus, the participants might develop a bias toward interpreting the initial portion of the carrier phrase as preceding a [p]- or [t]-initial target (e.g., *wel eens pot*). This would penalize a cluster-initial interpretation of the phrase (e.g., *wel een spot*), thus reducing transitory fixations to the competitor (e.g., *spin*). To prevent this, 25 filler trials were constructed that included (1) cluster-initial targets preceded by the carrier phrase (e.g., *ze heeft wel een sleutel gezegd* [she did say one key]), (2) targets starting with phonemes other than [p] or [t] preceded by the carrier phrase with word-final [s] (e.g., *zij heeft wel eens maan gezegd* [she said once moon]), (3) targets starting with [p] or [t] but preceded by the carrier phrase as it appears in the cluster-initial interpretation (e.g., *ze heeft wel een pauw gezien* [she did see a peacock]), (4) targets starting with phonemes other than [p] or [t] preceded by the carrier phrase as it appears in the cluster-initial interpretation (e.g., *ze heeft wel een bel geschreven* [she did write a bell]), and (5) targets starting with phonemes other than [p] or [t] preceded by a phrase very similar to the carrier phrase (e.g., *hij heeft wel vaker meloen gekocht* [he did buy melon often]). In addition, five filler items contained targets starting with [p] or [t] preceded by the carrier phrase (e.g., *zij heeft wel eens pak gezegd* [she said once suit]) but not causing any lexical ambiguity (i.e., *spak* is not a Dutch word). The other 20 filler items did not contain the carrier phrase. These sentences had diverse syntactic and prosodic structures (e.g., *zij probeerde een asbak te vinden* [she tried to find an ashtray]).

The sentences mentioning the filler items were produced by the same speaker and recorded at the same time as the experimental sentences. Cross-spliced sentences were constructed for 19 filler items containing the carrier phrase. The splicing procedure was similar to that carried out with the experimental items; that is, the filler word and the [s] preceding it were spliced from one token of the sentence onto another token. Three items (*bak* [bowl], *klok* [clock], *scepter* [scepter]) proved to be problematic to cross-splice without creating acoustic artifacts and had to be excluded from the experiment.

Acoustic analyses. Acoustic measurements of the ambiguous sequences (e.g., *eens pot* vs. *een spot*) were carried out to evaluate the extent to which the meaning intended by the speaker influenced the way the sequences were produced. The following durational measurements were made: the duration of the segments [ə], [n], and [s], the duration of the closure (before the stop), voice onset time (VOT) of the stop, and the duration of the word excluding the stop. These measurements were based on an analysis of both spectrograms and waveforms. RMS energy and spectral center of gravity (SCG) were measured for the [s] and for the stops. RMS energy was calculated by taking the logarithm of the root mean sum of squares of all sample points in the segment. SCG of stops was measured using the built-in function in the Praat speech editor (www.praat.org). This function calculates the average frequency from an FFT spectrum over a frequency range from 0 to 10000 Hz. The SCG of [s] was measured by dividing the segment into 15-msec intervals, computing an FFT spectrum for each interval (filtering out frequencies below 1000 Hz in order to remove any spurious low-frequency components) and taking the SCG of each interval. The SCG of the segment was the maximal SCG value across all intervals.

Procedure and Design. The participants were tested individually. To ensure that they identified the pictures as intended, the participants were first familiarized with all 268 pictures. The pictures appeared on a computer screen in random order, one at a time, along with their printed names. The participants were instructed to familiarize themselves with each picture and then to press a button to go on to the next picture. The eye-tracking system was then set up.

The participants were seated in front of the computer screen at a comfortable viewing distance. The eye-tracking system was mounted and calibrated. Eye movements were monitored using an SMI EyeLink eye-tracking system, sampling at 250 Hz. The experiment was controlled by a Compaq 486 computer. Pictures were presented on a ViewSonic 17PS screen. Auditory stimuli were pre-

Table 1
Stimulus Example of the Conditions in Experiment 1

Origin of Recording	Example	Spliced Version
	Identity-Spliced Condition	
Target Context 1	Ze heeft wel eens pot gezegd	Ze heeft wel een*s pot* gezegd
Target Context 2	*Ze heeft wel eens pot gezegd*	
	Cross-Spliced Condition	
Target Context 1	Ze heeft wel eens pot gezegd	Ze heeft wel eens pot gezegd
Cluster-Initial Context	Ze heeft wel een spot gezegd	

sented over headphones using NESU software (www.mpi.nl/world/tg/experiments/nesu.html). Both eyes were monitored, but only the data from the right eye were analyzed.

Each trial had the following structure. A central fixation dot appeared on the computer screen for 500 msec. A spoken sentence was then presented and, simultaneously, a 5 × 5 grid with pictures appeared on the screen (see Figure 1). The participants had received written instructions to move the object mentioned in the spoken sentence above or below the geometrical shape adjacent to it, using the computer mouse. The positions of the pictures were randomized over trials across the four fixed positions of the grid shown in Figure 1, but the geometric shapes always appeared in fixed positions. The participants' fixations for the entire trial were completely unconstrained, and they were put under no time pressure to perform the action. After the participant had moved the picture, the experimenter initiated the next trial. The software controlling stimulus presentation (pictures and spoken sentences) interacted with the eyetracker output so that the timing of critical events in the course of a trial (such as the onsets of the spoken stimuli and mouse movements) was added to the stream of continuously sampled eye-position data. After every five trials, a fixation point appeared centered on the screen, and the participants were instructed to look at it. The experimenter could then correct potential drifts in the calibration of the eyetracker.

Two lists were created, each containing 47 filler items and 20 experimental items. Within each list, 10 experimental items were assigned to the identity-spliced condition and 10 to the cross-spliced condition. The only difference between the two lists was thus which version of each experimental sentence was presented. The participants were randomly assigned to one list. Twelve random orders of presentation were created, with the constraints that there was always at least one filler item between two experimental items and that five of the filler trials were presented at the beginning of the experiment to familiarize the participants with the task and procedure.

Results and Discussion

The data from each participant's right eye were analyzed and coded in terms of fixations, saccades, and blinks, using the algorithm provided in the EyeLink software. Timing of fixations was established relative to the onset of the critical [s] (i.e., the splice point) in the spoken utterance. Graphical software displayed the locations of the participants' fixations as dots superimposed on the four pictures used in each trial. The fixation dots were numbered in the order in which the fixations occurred. Fixations were coded as pertaining to the target, to the competitor, to one of the two unrelated distractors, or to anywhere else on the screen. Fixations that fell within the cell of the grid in which a picture was presented were coded as fixations to that picture. For each experimental trial, fixations were coded from the splice point (the onset of the preceding [s]) until the participants had clicked on the target picture with the mouse, which was taken to reflect the participants'

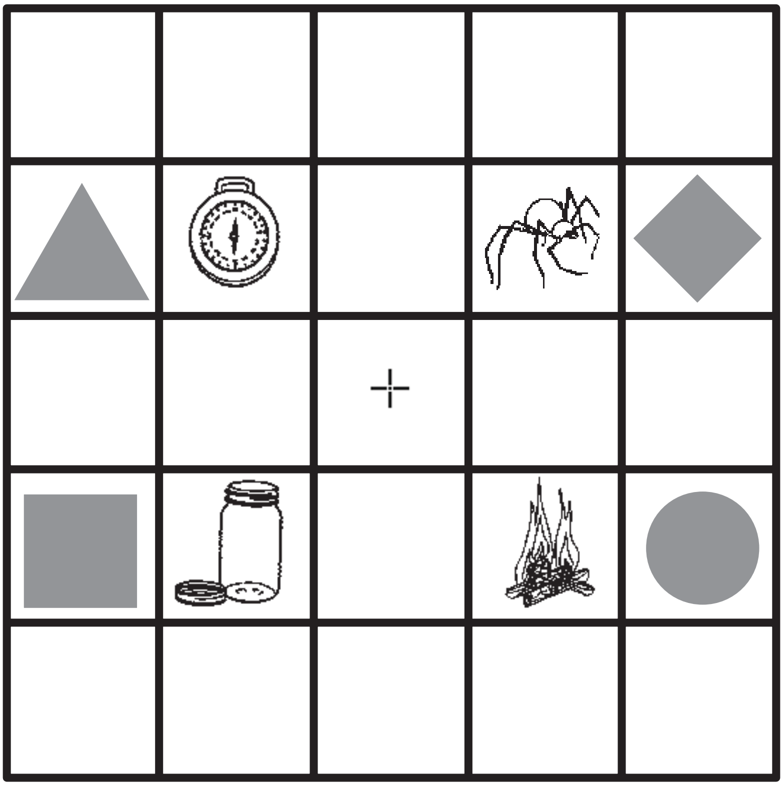

Figure 1. Example of stimulus display presented to participants. The geometric shapes (triangle, diamond, circle, and square) and the central fixation cross were in fixed positions across trials. The pictured objects and their positions varied over trials. In this example, these were, clockwise from top left: *kompas* (compass), *spin* (spider), *vuur* (fire), and *pot* (jar).

identification of the referent. In most cases, the participants were fixating the target picture when clicking on it. In the rare cases in which the participants clicked on the target picture while not simultaneously looking at it, an earlier fixation to the target picture was taken as indicating recognition of the target word, and the coding of the trial ended with that fixation. On two trials, the participants erroneously moved an object other than the target picture without correcting their choice. These trials were excluded from the analyses.

For each participant, fixation proportions were averaged across items, separately for each condition. The proportions of fixations to each picture or location (i.e., target picture, competitor picture, distractor pictures, or elsewhere) were computed for each 10-msec slice. Blinks and saccades were not included in this calculation. A similar analysis was done for each item, averaging across participants.

Figure 2 presents the proportions of fixations averaged over participants in the identity-spliced (Figure 2A) and cross-spliced (Figure 2B) conditions. Fixation proportions for the two unrelated distractors in each condition were averaged. In Figure 3, the proportions of fixations to the targets and competitors in both splicing conditions are displayed. All figures show fixation proportions in 10-msec time slices from the splice point (the onset of the [s] preceding the target word) to 1,200 msec thereafter.

As is apparent from Figure 2, fixation proportions to the competitor pictures began to rise in both conditions at around 300 msec. It is estimated that an eye movement is typically programmed about 200 msec before it is launched (e.g., Fischer, 1992; Hallett, 1986; Matin, Shao,

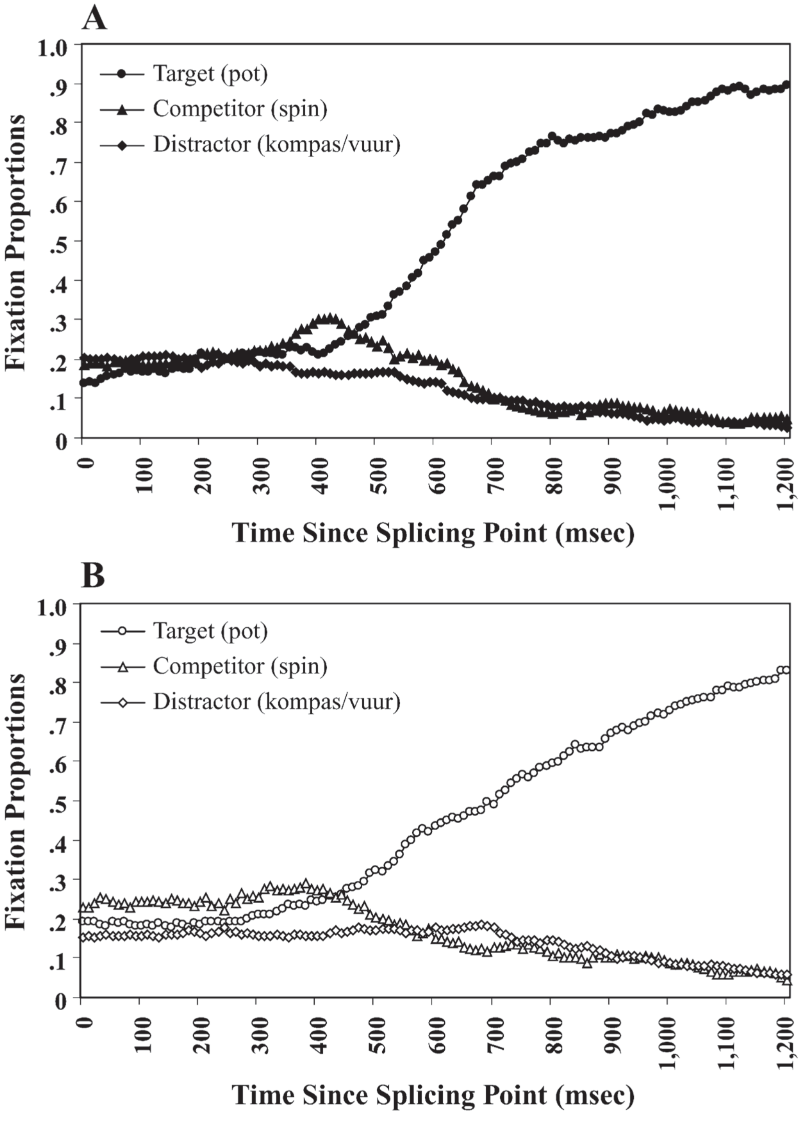

Figure 2. Fixation proportions over time to the target, the competitor, and averaged distractors in the identity-spliced condition (A) and the cross-spliced condition (B) in Experiment 1.

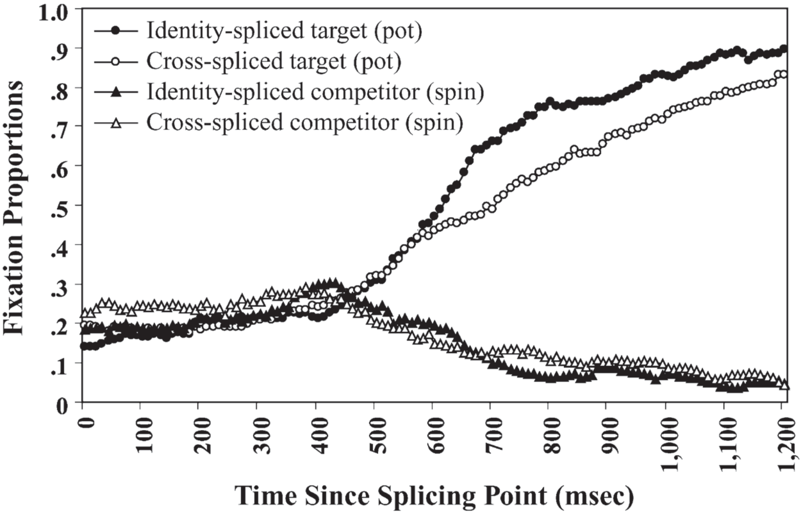

Figure 3. Fixation proportions over time to the target and competitor pictures in the identity-spliced and the cross-spliced conditions in Experiment 1.

& Boff, 1993; Saslow, 1967), so that 300 msec after the splicing point is approximately the moment at which fixations driven by the first 100 msec of acoustic information after the splicing point can be seen. Thus, the mapping of the acoustic signal onto the lexical representations is reflected by fixations from about 300 msec on.[2] In both conditions, fixation proportions to the competitor pictures remained higher than those to the distractor pictures until around 600 msec, where they merged again.

Fixation proportions to the target pictures also began to increase around 300 msec, in both conditions, and rose above fixation proportions to the competitor at around 450 msec. The fixation proportions in the two conditions increased initially with a similar slope, but at around 600 msec, the proportions of fixations started to diverge, with fixation proportions to the target in the identity-spliced condition rising faster and remaining higher than those to the target in the cross-spliced condition.

The difference between conditions was statistically tested by computing the average fixation proportion to the target picture over a time window extending from 300 to 1,200 msec. Over this time interval, the average fixation proportion to the target picture was .60 in the identity-spliced condition and .53 in the cross-spliced condition. A one-factor ANOVA on the mean proportion of fixations was conducted over this time window, with splicing (identity-spliced condition vs. cross-spliced condition) as a within-participants factor. In the item analysis, splicing was a between-items factor.[3] Targets in the identity-spliced condition were fixated significantly more often than targets in the cross-spliced condition [$F_1(1,23) = 13.99, p < .01, \eta^2 = .38$; $F_2(1,19) = 6.95, p < .05, \eta^2 = .27$]. Additionally, fixations to the competitor and distractor pictures were compared over the time window extending from 300 to 600 msec. Over this time period, there were more fixations to the competitors than to the distractors (.24 and .16, respectively). In a two-way (picture [competitor vs. distractor] × splicing condition) ANOVA, this difference was significant [$F_1(1,23) = 13.81, p = .001, \eta^2 = .36$; $F_2(1,19) = 19.04, p < .001, \eta^2 = .5$], but there was no effect of splicing (average fixation proportions were .21 and .20 in the identity- and cross-spliced conditions, respectively; F_1 and $F_2 < 1$). Furthermore, the interaction between the factors was not significant, indicating that fixation proportions to the competitor pictures were equally high in both conditions.

The eye-tracking results demonstrate that the sequences presented in the two conditions, though phonemically identical, contained fine-grained differences that the participants were sensitive to, resulting in modulation of their lexical interpretation. To examine these fine-grained differences, acoustic analyses were conducted on the two ambiguous sequences. The results of the acoustic measurements are displayed in Table 2. The results of one-way ANOVAs performed on these data are presented in the same table. These analyses revealed significant differences between the two sequences on several measures: (1) duration of the [s] in the target word context was shorter than that in the cluster-initial word context, (2) closure duration was longer in the target context than in the cluster-initial context, (3) duration of the target words (measured from after the release of the stop) was longer than duration of the cluster-initial words, (4) RMS energy of [s] in the target context was lower than in the cluster-initial context, and (5) RMS energy of the stop in the target words was lower than in the cluster-initial words.

The acoustic measurements showed that there were subtle differences between the two versions of the ambiguous sequences. These acoustic differences between the target and cluster-initial sequences are effective cues, however, only to the extent that listeners can perceive these differences and use them in word recognition. For

Table 2
Mean Segmental Duration (in Milliseconds), RMS Energy (in Decibels), Spectral Center of Gravity (SCG, in Hertz), and Standard Deviations (*SD*s) of the Ambiguous Sequences in the Experimental Sentences

	Target Word *eens pot*		Cluster-Initial Word *een spot*		ANOVA	
	M	*SD*	*M*	*SD*	*F*(1,19)	*p*
Duration						
[ə]	55	8	55	10	<1	n.s.
[n]	22	4	20	5	2.05	n.s.
[s]	91	15	108	14	19.72	<.001
Closure	81	25	59	22	39.57	<.001
Voice onset time	22	8	22	7	<1	n.s.
Word (excluding stop)	193	46	181	43	9.34	<.01
RMS Energy						
[s]	3.28	.11	3.37	.11	6.85	<.05
stop	3.11	.15	3.21	.13	4.10	=.057
SCG						
[s]	5,322	372	5,458	392	1.73	n.s.
stop	1,231	995	1,487	1,207	3.23	n.s.

the acoustic measurements for which a significant difference was found, the difference in the measurements for each item was therefore correlated with the perceptual effect for that item (i.e., the difference in average fixation proportions to the item between the identity-spliced and the cross-spliced conditions in the time window extending from 300 to 1,200 msec).

As shown in Table 3, there were no significant correlations in the initial analysis between any of the acoustic measurements and the perceptual effect. However, scatter plots of the difference in acoustic measurements against the perceptual effect indicated that, for the duration of the [s], the lack of correlation was caused by the presence of one outlier in the data set (see Figure 4). When this outlier (the item *thee* [tea]) was removed, a strong linear correlation emerged [$r(19) = .60, p < .01$]. The exclusion of the outlier did not improve the correlation of the other measurements with the perceptual effect. There were no such outliers for any of the other measurements. Furthermore, when the differences in the acoustic measurements were entered into a stepwise linear regression analysis, only the difference in the duration of the [s] was found to be a significant predictor of the perceptual effect, accounting for 32% of the variance (adjusted $r^2 = .32$). Thus, the data suggest that, although the ambiguous sequences differed on several measurements, the participants used only the duration of the [s] as a cue for the word boundary.

Table 3
Correlation of the Difference in the Acoustic Measurements With the Perceptual Effect

Measurement	*r*(20)	*r*(19)*
Duration of [s]	.263	.600
Closure duration	.184	.142
Word duration (excluding stop)	.243	.148
RMS energy of [s]	.381	.261
RMS energy of stop	.176	.237

*Outlier excluded.

One interesting aspect of the data concerns the timing of the splicing effect. As is apparent in Figure 3, the difference between the identity-spliced and the cross-spliced conditions started to take place around 600 msec after the splicing point. Indeed, statistical analyses across the 300- to 1,200-msec time frame in 100-msec bins indicated that fixations to the target in the identity-spliced condition started differing significantly from the fixations to the target in the cross-spliced condition in the 600- to 700-msec time bin [$F_1(1,23) = 12.78, p < .01$; $F_2(1,19) = 6.93, p < .05$]. If one assumes a 200-msec delay in programming and launching an eye movement, this would mean that the difference started to appear after 400 msec of the postsplice portion of the ambiguous sequence had been heard. Given that the spliced portion of the stimulus was, on average, 380 msec long, this indicates that the effect started to take place around word offset (i.e., at the end of the spliced portion). Considering that the information that seems to be most important for the effect (i.e., the duration of the [s]) occurs early in this portion, one might have expected the effect to start earlier. Instead, the data suggest either that additional information about the ambiguous sequence needs to accumulate or that more processing time is required before the duration of the [s] starts to bias the sequence's interpretation. This implies that the duration of the [s] alone may not be able to cause an immediate effect. We examined this issue in Experiment 2 by manipulating the duration of the [s].

Another motivation for running Experiment 2 was that the results of Experiment 1 could perhaps be attributed to the splicing manipulation we performed. It could be the case that cross-spliced stimuli elicited fewer fixations to the target not because of the specific acoustic-phonetic information they contained (i.e., the duration of the [s]) but due to some nonspecific acoustic factors that caused them to be of poorer quality. Although the correlation of [s] duration with the behavioral effect suggests that this is not the case, we addressed this concern directly in Experiment 2. The

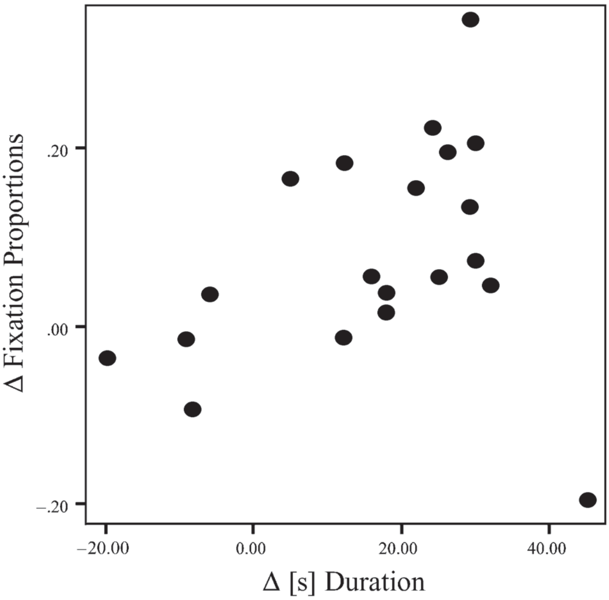

Figure 4. Scatter plot of the difference between the identity-spliced and cross-spliced conditions of Experiment 1 in fixation proportions to the target against the difference between the conditions in the duration of the [s].

same physical sentence was used in both conditions, with the duration of the [s] either shortened or lengthened.

EXPERIMENT 2

The purpose of Experiment 2 was to evaluate the degree to which the duration of the [s] in an ambiguous sequence (such as *eens pot*) can bias its lexical interpretation, when the acoustic information in the rest of the sequence is held constant. The results of Experiment 1, and in particular the timing of the effect, suggest that durational information was not evaluated on its own, but rather relative to other accumulating information (either from the signal or from the processing thereof). In Experiment 2, we examined whether this would also be the case when the duration of the [s] would render it very likely to be in either word-final or word-initial position. To do this, the distribution of [s] duration in word-final and word-initial positions in our original recordings was taken into account. The values for [s] duration in Experiment 2 were chosen such that the [s] in one condition (the short version) fell clearly within the distribution of word-final [s], and the [s] in the other condition (the long version) fell clearly within the distribution of word-initial [s]. The stimuli were created by shortening or lengthening the duration of the [s] such that it was either 1 standard deviation (*SD*) below the mean of the word-final distribution (short version) or 1 *SD* above the mean of the word-initial distribution (long version). We predicted that there would be fewer fixations to the target in the long-version condition. By using the same stimuli that were used in Experiment 1, we hoped to be able to make a comparison between the two experiments regarding the time course of the effect.

Method

Participants. Twenty-four student volunteers from the Max Planck Institute subject pool were paid for their participation. They were all native speakers of Dutch. None of them had participated in the previous experiment.

Materials. New stimuli were created by manipulating the duration of the [s] consonant in the target context sentences from our original recording (e.g., the unspliced sentence *ze heeft wel eens pot gezegd*). For each sentence, two spliced versions were created, in which the duration of the [s] was either shortened or lengthened. In determining which value the duration of the [s] in the edited versions should take, we examined the distribution of [s] durations in the original recording. Over all the tokens, the duration of the [s] was 87 msec (*SD* = 15) when it was in word-final position, and 107 msec (*SD* = 14) when it was in word-initial position. On the basis of these numbers, for the version with the short [s] duration, the duration of the [s] was selected to be approximately 1 *SD* lower than the mean duration of the [s] in word-final position, resulting in a value of 70 msec. For the long version, the duration of the [s] was approximately 1 *SD* higher than the mean duration of the [s] in word-initial position, resulting in a value of 121 msec. The [s] durations were thus relatively extreme, given the distribution in the original recording, but still well within this speaker's normal range.

The stimuli were edited using the Xwaves speech-editing software. Durations of the [s] were manipulated by cross-splicing. In each sentence, the steady-state phase of the fricative was excised, leaving approximately 20 msec of the initial and final portions of the frication noise (subject to small variation due to the restriction of splicing at zero crossings). The steady-state phase was replaced by a fragment of steady-state [s] frication (from another token), which was either 30 msec long or 80 msec long, resulting in fricatives that had durations of, respectively, 70 msec (short version) or 120 msec (long version). Care was taken to avoid any acoustic artifacts, such as clicks or other distortions.

Procedure and Design. The procedure and design were identical to those of Experiment 1.

Results and Discussion

On two trials, the participants erroneously moved an object other than the target picture without correcting their choice. These trials were excluded from the analyses. Figure 5 presents the proportions of fixations averaged over participants in the short-version condition (Figure 5A) and the long-version condition (Figure 5B). Fixation proportions for the two unrelated distractors were averaged. In Figure 6, the proportions of fixations to the targets and competitors in both duration conditions are displayed.

Figure 5A shows that the probability of fixating the competitor in the short-version condition began to diverge from the probability of fixating the unrelated distractors about 200 msec after the splicing point. At the same time, fixations to the target picture were also rising. At around 450 msec, fixations to the target rose above those to the competitor. From that point on, the probability of fixating the competitor started to drop, and at around 600 msec it was indistinguishable from the probability of fixating a distractor.

Figure 5B shows a somewhat different pattern of fixation proportions in the long-version condition. Fixations

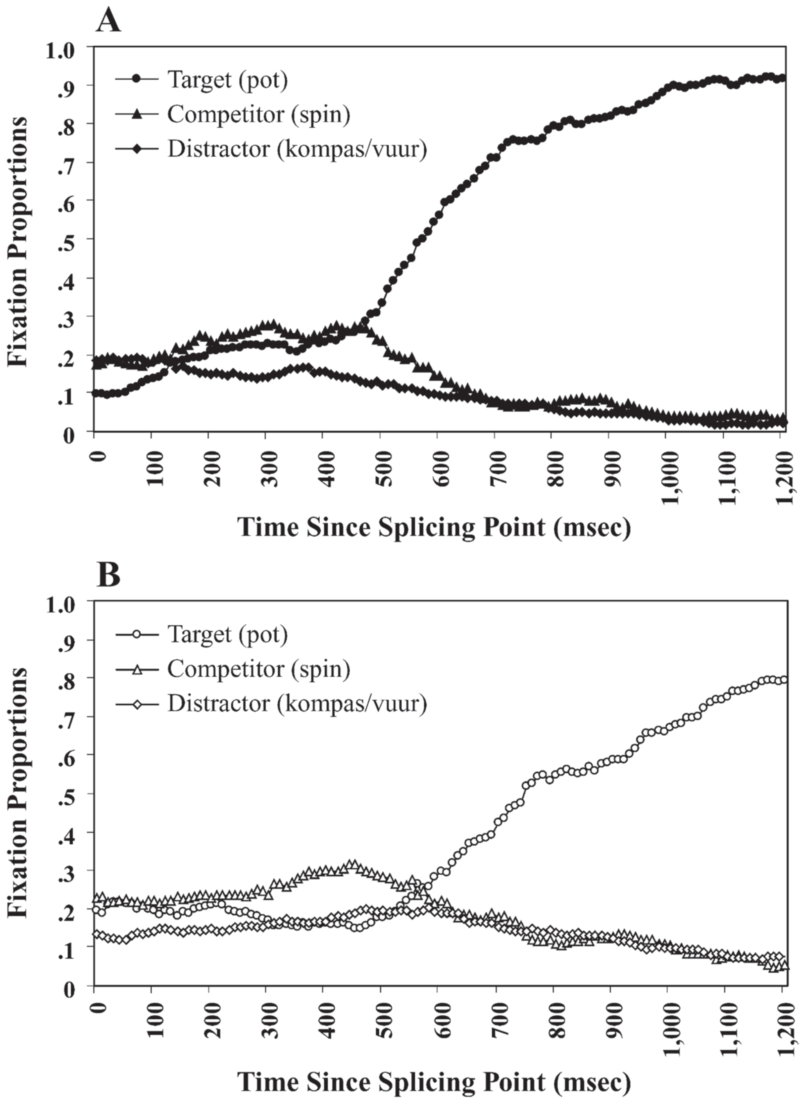

Figure 5. Fixation proportions over time to the target, the competitor, and averaged distractors in the short-version condition (A) and the long-version condition (B) in Experiment 2.

to the competitor gradually rose and reached a peak at about 450 msec after the splicing point, and the probability of fixating the target picture remained indistinguishable from the probability of fixating the unrelated distractors. At around 450 msec, fixations to the target picture started to increase while fixations to the competitor were decreasing, until around 600 msec, where they merged again with the fixation proportions of the unrelated distractors.

Figure 6 presents the proportions of fixations over time to the target and competitor pictures, in both conditions. As is immediately apparent from the graph, there was a major effect of condition, such that, from as early on as 250 msec after the splicing point, the participants tended to fixate the target picture less when they heard the long [s] version of the sentence. Over the 300- to 1,200-msec time window, the average proportion of fixations to the target picture was .64 in the short-version condition and .45 in the long-version condition. A one-way ANOVA showed that this effect was statistically significant [$F_1(1,23) = 91.55$, $p < .001$, $\eta^2 = .80$; $F_2(1,19) = 44.17$, $p < .001$, $\eta^2 = .70$]. Given that the duration of the [s] was longer in one condition, it could be argued that the effect was partly due to the delay in the onset of the target in the signal. The data were therefore realigned to the point of the onset of the stop closure (the offset of the [s]). Analysis of the realigned data showed that the pattern of the results remained unchanged. Over the 300- to 1,200-msec time window, the average proportion of fixations to the target picture was .70 in the short-version condition and .54 in

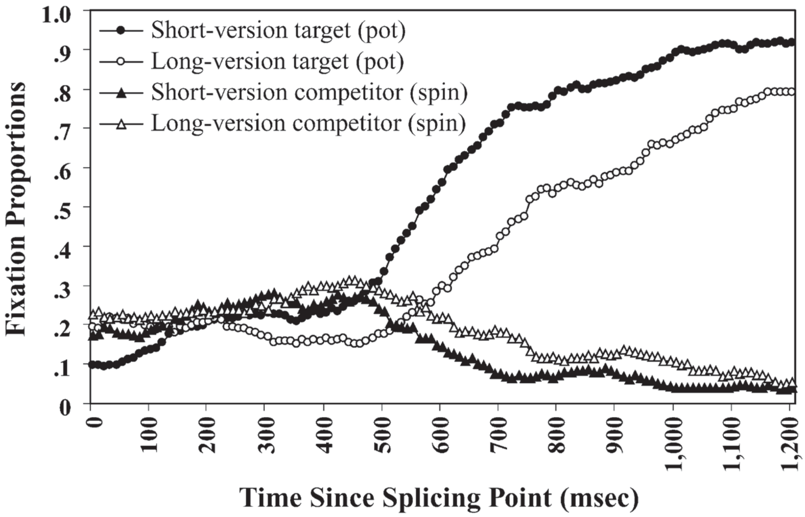

Figure 6. Fixation proportions over time to the target and competitor pictures in the short-version and the long-version conditions in Experiment 2.

the long-version condition, yielding a significant effect [$F_1(1,23) = 50.81, p < .001, \eta^2 = .69; F_2(1,19) = 23.68, p < .001, \eta^2 = .55$].

Fixations to the competitor and distractor pictures were compared over the time window extending from 300 to 600 msec. Over this time period, there were more fixations to the competitors than to the distractors (.26 and .16, respectively). In a two-way (picture × splicing condition) ANOVA, this difference was significant [$F_1(1,23) = 16.54, p < .001, \eta^2 = .42; F_2(1,19) = 19.21, p < .001, \eta^2 = .5$]. There was also a significant effect of splicing, such that pictures in the long-version condition were fixated more than pictures in the short-version condition [average fixation proportions were .23 and .18 in the short and long-version conditions, respectively; $F_1(1,23) = 5.89, p < .05, \eta^2 = .20; F_2(1,19) = 24.81, p < .001, \eta^2 = .57$]. However, the interaction between the factors was not significant, indicating a statistically equivalent effect of the splicing manipulation on the competitor and the distractors. In other words, in the long [s] condition, the participants looked more often at the competitors (relative to in the short [s] condition), but they also looked more often at the distractors. These results thus do not indicate that the long [s] version was a better match for the competitor (relative to the short [s] version). Rather, it appears that the long [s] version was a poorer match for the target. Given that fixation proportions to the different pictures are not independent of each other, lower fixation proportions for the target in this experiment entail higher fixation proportions for all the other pictures. It should be noted, however, that the absence of a splicing effect specifically on the fixations to the competitor does not in any way count against the conclusion that segment duration guides segmentation. The overlap of the competitor with the signal is rather small (just the cluster). This means that the period during which the competitor is a likely candidate is short. Consequently, there is little time for a splicing effect on the competitors to take place. Furthermore, in the long [s] condition, words that start with an [s] (and followed by a [p] or a [t]) are favored, but the signal does not provide any additional support for the competitor itself to be a favored candidate (among the cohort of cluster-initial words).

Similarly to what was observed in Experiment 1, the participants in Experiment 2 were slower to fixate the target when the duration of the [s] in the ambiguous sequence was long. In contrast to Experiment 1, in which the effect emerged only late in the trials, in Experiment 2 the effect of the splicing manipulation appeared almost as soon as the disambiguating information was heard. Statistical analyses across the 300- to 1,200-msec time frame in 100-msec bins confirmed this difference in the timing of the effect. In the 300- to 400-msec time bin, fixations to the target in the short-version condition started differing from the fixations to the target in the long-version condition, a difference that was reliable in the participants analysis [$F_1(1,23) = 5.07, p < .05, \eta^2 = .18$], though not in the items analysis [$F_2(1,19) = 2.45, p = .13$]. In the 400- to 500-msec time bin, this difference was significant [$F_1(1,23) = 12.32, p < .01, \eta^2 = .35; F_2(1,19) = 11.61, p < .01, \eta^2 = .38$]. The difference between conditions thus arose earlier in Experiment 2 than in Experiment 1.

A two-way (condition × experiment) ANOVA on fixation proportions to the target over the 300- to 1,200-msec interval was then conducted to compare directly the results of the two experiments. Experiment was treated as a between-participants factor in the F_1 analysis and as a within-items factor in the F_2 analysis. The analysis revealed a significant effect of condition [$F_1(1,46) = 88.79, p < .001, \eta^2 = .66; F_2(1,19) = 28.28, p < .001, \eta^2 = .6$], no main effect of experiment, and, critically, a significant interaction between condition and experiment

[$F_1(1,46) = 17.21, p < .001, \eta^2 = .27; F_2(1,19) = 14.20, p < .01, \eta^2 = .43$]. This analysis indicates that, although in Experiment 2 only the duration of the [s] was manipulated, the fact that the values taken for [s] duration were relatively extreme caused the behavioral effect to be significantly larger than in Experiment 1.

The results of Experiment 2 confirm that the duration of the [s] can modulate the interpretation of an ambiguous sequence. Moreover, the time course of the effect in Experiment 2 demonstrates that, if the duration of the [s] indicates clearly which position the [s] is likely to appear in, the perceptual system can use this information very quickly to bias the interpretation of the sequence, without requiring additional time or additional information. Thus, segment duration on its own can bias participants' interpretation of lexically ambiguous sequences.

GENERAL DISCUSSION

Dutch listeners use the duration of individual speech sounds as a cue to the location of word boundaries in their online segmentation of continuous speech. The participants listened to sentences in which a stop-initial target word (e.g., *pot*) was preceded by an [s], thus causing ambiguity regarding whether the sentence referred to a stop-initial word or a cluster-initial word (e.g., *spot*). The participants' fixations to a picture representing the target word (e.g., a jar) were taken to reflect the degree of lexical activation of that word. In Experiment 1, the participants were slower to fixate the target pictures when the sentences were manipulated such that the target and the preceding [s] were spliced from a recording of the cluster-initial word than when the target and the preceding [s] were spliced from a different token of the sentence containing the stop-initial word. Acoustic analyses showed that the two versions differed in various measures, but only one of these (the duration of the [s]) correlated with the perceptual effect. In Experiment 2, the sentences containing the target words were manipulated such that the duration of the [s] preceding the target was either lengthened or shortened. The participants were slower to fixate the target pictures when the duration of the [s] was lengthened than when it was shortened. Taken together, these results demonstrate that, in the context of these ambiguous sequences, the duration of the [s] is an important determinant of the lexical interpretation of this type of utterance.

Similar results have been obtained in another eye-tracking study (Shatzman, 2004). This experiment was a variant of the present study: It used the same sentences but a different splicing manipulation. The initial stop of the target word and the preceding [s] (e.g., the [s] and the [p] in *eens pot*) were replaced either by a cluster from the cluster-initial word (e.g., the [sp] from *spot*) or by an initial stop and preceding [s] from another recording of the sentence. The participants made fewer fixations to the target pictures when the stop and the preceding [s] were cross-spliced from the cluster-initial word than when they were spliced from the sentence containing the stop-initial word. As in the present study, acoustic analyses showed that the two versions differed in various measures, but only the duration of the [s] correlated with the perceptual effect.

These results are consistent with previous findings (Kemps, 2004; Quené, 1992) that have demonstrated in offline tasks that listeners use phoneme duration to segment ambiguous sequences. Using the eye-tracking paradigm, the present study extends those findings by showing that listeners use phoneme duration in the online segmentation of ambiguous phrases. Furthermore, unlike previous studies of online segmentation (e.g., Gow & Gordon, 1995; Spinelli et al., 2003), in which it was assumed that segment duration differences found between the materials were used by listeners to disambiguate the phrases, the present study has shown that the perceptual effect correlated with the duration of the [s] and that manipulating the [s] duration alone can bias participants' interpretation of the ambiguous sequence. Thus, our study provides evidence that directly links individual segment duration and listeners' lexical interpretation.

Moreover, our study has shown that finding an acoustic difference between the two recording contexts of the ambiguous phrases (i.e., *eens pot* vs. *een spot*) does not necessarily mean that listeners will attend to that difference. In addition to the difference in [s] duration, the two recording contexts differed in the duration of the closure before the stop, the duration of the target word (excluding the stop), RMS energy of the [s], and RMS energy of the stop. Any of these measurements could potentially be used as a cue to differentiate the two possible readings of the ambiguous phrase. Our correlational analysis showed, however, that segmentation was not influenced by these other differences, but rather that listeners were relying on the duration of the [s]. That is not to say that the other acoustic measurements cannot influence segmentation. It is possible that manipulating one of these measures, while keeping [s] duration constant, would affect segmentation. The results of the present study indicate, however, that, given normal variation in natural speech, listeners' segmentation of ambiguous sequences such as *eens pot*/*een spot* can be best predicted from the duration of the [s].

As noted earlier, there is considerable variation among speakers in how prosodic boundaries are realized (e.g., Fougeron & Keating, 1997). In the present study, it has been assumed that the acoustic measurements of the speaker's utterances can be generalized to other speakers, though we have not carried out a larger production study to confirm whether this is indeed the case. With regard to the duration of the [s], however, our results can be compared with the results of the study of Waals (1999), in which the duration of various consonants in Dutch was measured, in various word positions. In Waals's study, the duration of [s] in a word-initial cluster (followed by a [t] or a [p]) was 107 msec, whereas a word-final [s] (in a cluster, preceded by [n], as in the word *eens*) was on average 76 msec. Hence, the duration of word-initial [s] in

Waals's study was virtually identical to that of the speaker used for the recording of our materials (108 msec). In our study, word-final [s] duration tended to be longer than that in Waals's study (91 msec in the stimuli used in Experiment 1 and 87 msec over all tokens). It therefore seems that, in our speaker's speech, the difference between word-initial and word-final [s] was slightly smaller than that found in Waals's study. We may therefore have underestimated the acoustic difference between the two types of utterances. Although the strong similarities between the two studies suggest that our findings are generalizable over speakers, we cannot yet be confident that this is the case. What our study does show, however, is that if the speaker exhibits a contrastive durational pattern for word-initial and word-final [s], listeners will exploit this information in segmentation. Moreover, we cannot predict with any confidence whether the present findings will generalize to segments other than [s]. It seems quite reasonable, in fact, to assume that with different segments, as well as in different languages, other acoustic cues might be used by listeners in lexical disambiguation. It therefore remains an open question whether segment duration is always the most important of these cues.

The findings of the present study add to a growing body of research showing that fine-grained information in the speech signal can modulate lexical activation (Andruski, Blumstein, & Burton, 1994; Dahan et al., 2001; Davis et al., 2002; Gow, 2002; Marslen-Wilson & Warren, 1994; McQueen, Norris, & Cutler, 1999; Salverda et al., 2003; Streeter & Nigro, 1979; Tabossi, Collina, Mazzetti, & Zoppello, 2000). The picture emerging from these studies, and from the present results, is that the speech-recognition system is able to pick up subtle acoustic differences in the speech signal and use this information to modulate lexical activation in favor of the intended word.

One way in which models of spoken-word recognition could accommodate these findings is to assume that durational information is part of stored lexical knowledge. On this account, durational differences are viewed as inherent properties of lexical representations, to which the incoming signal is directly compared. In such exemplar-based models (e.g., Goldinger, 1998; Johnson, 1997a, 1997b), stored exemplars of words with an [s] in word-final position (such as the Dutch word *eens*) would have shorter [s] duration than stored exemplars with an [s] in word-initial position (e.g., *spot*); therefore, an ambiguous phrase such as *eens pot* with a long [s] duration would bias the system to interpret the sequence as containing an [s] in word-initial position.

Another way in which these findings can be modeled is to have durational information bias prelexical representations. Models such as TRACE (McClelland & Elman, 1986), Shortlist (Norris, 1994), and the DCM (Gaskell & Marslen-Wilson, 1997) incorporate prelexical representations that recode the speech signal in some abstract way prior to lexical access. It is clear, however, that phonemic prelexical representations cannot provide an adequate account of the available data on sensitivity to fine-grained acoustic detail (see McQueen, Dahan, & Cutler, 2003, for discussion). Position-specific segmental representations at the prelexical level could explain the present results, however. Consonants of long duration could activate syllable-initial allophones more strongly than consonants of shorter duration, and consonants of short duration could activate syllable-final allophones more strongly than consonants of longer duration. A long [s], for example, could thus provide more support for the *een spot* reading, whereas a shorter [s] could preferentially activate the *eens pot* reading.

A third way in which to model the modulation of lexical activation by durational information is to assume that, in parallel to the segmental analysis of the utterance, a suprasegmental analysis is carried out in which a prosodic structure is built. According to this proposal (cf. Cho, McQueen, & Cox, in press; Salverda et al., 2003), durational information is used to signal the location of likely prosodic boundaries equal to or higher than the word. On this account, lexical candidates whose word boundaries are aligned with the predicted prosodic boundary are favored. Thus, for example, a short [s] would suggest a likely upcoming word boundary, resulting in a prosodic structure consistent with that of *eens pot* but not of *een spot*.

The data presented here are insufficient to distinguish among these three alternative accounts. We did, however, obtain results that impose important constraints on the accounts offered by all three of these models. The results of Experiment 1—and, more specifically, the time course of the effect—suggest that durational information is not evaluated on its own but rather relative to other accumulating information. Given the variability of speech, it seems very plausible that this would be the case. That is, it is unlikely that the speech recognition system would use absolute segment duration, because the same absolute duration can be long in one context (e.g., for one speaker at a given speaking rate) but short in another.

From the time course of the effect in Experiment 2, it transpires that, under certain circumstances, segment duration can bias the interpretation of the ambiguous sequence almost immediately. This was the case when the duration of the phonetic segment indicated that it is very likely to appear in a certain position in the word. The difference in the time course between the two experiments therefore suggests that durational information is used in a probabilistic way by the speech-recognition system. An undoubtedly long segment duration (as in Experiment 2) makes the probability that the phoneme is in word-final position low; therefore, such an interpretation is disfavored. If, however, segment duration does not clearly indicate which position the phoneme is likely to occupy (as was the case in Experiment 1), it seems that more information needs to accrue for the system to determine which position is most probable for the phoneme and, consequently, which interpretation of the ambiguous sequence is favored.

The fact that segment duration seems to be evaluated in a relativistic manner is not surprising, given this cue's temporal nature and the variability that exists in the temporal structure of speech. Previous studies have identified several temporal cues that are perceived in relation to speech rate (e.g., Miller & Liberman, 1979; see Miller, 1981, for a review). For example, listeners' ratings of the goodness of stimuli as instances of a phonetic category depend on speaking rate (e.g., Allen & Miller, 2001; Miller & Volaitis, 1989). Independent of speaking rate variation, fricative duration (at least in English) is also used in fricative identification (e.g., Cole & Cooper, 1975; Jongman, 1989; Stevens, Blumstein, Glicksman, Burton, & Kurowski, 1992). For example, in the study by Stevens et al., listeners appeared to use the length of an [s] as a cue to the voicing contrast between [s] and [z].

The results of the present study indicate, however, that durational cues can do more than affect the perception of an input segment in relation to contrasts between or within phonetic categories. In our experiments, varying the duration of the [s] did not matter for a phonetic distinction (the [s] was just as much an [s] in *eens pot* as in *een spot*). It did, however, influence the likelihood of the [s] to be in one word position or another. That is, varying [s] duration did not change the perceived goodness of the phoneme as that phoneme, but rather the perceived goodness of that phoneme as occurring in a certain position in the word. A clearly long [s] duration (i.e., long relative to the information that has accumulated up to that point), while still being perceived as a good [s], was apparently perceived as a poor exemplar of a word-final [s] and therefore biased the system toward one lexical interpretation.

The evaluation of durational differences as a function of word position is likely to be orthogonal to the evaluation of differences that are due to speaking rate and fricative category, however. As we have just argued, durational cues are evaluated relative to speaking rate in order to modulate segmental interpretation (see, e.g., Allen & Miller, 2001). Duration, independent of speaking rate, can also be used in segmental interpretation (see, e.g., Stevens et al., 1992). These processes of segmental evaluation could occur prelexically. But our results suggest that durational differences must also be able to modulate lexical-level processes. Note also that segment duration that does not unequivocally point to one probable position for the phoneme (given the durational information acquired up to that point) will require additional information in order to favor one lexical interpretation substantially over another. The challenge for any model of spoken-word recognition, therefore, is twofold. First, a mechanism must be developed that can evaluate fine-grained durational differences in both a relativistic and a probabilistic manner. Second, this mechanism must be able to use durational information in this manner both for segmental distinctions and for lexical distinctions that do not depend on differences between phonemes.

In the experiments reported in this article, we investigated the degree to which listeners use various acoustic cues to word boundaries, using lexically ambiguous phrases such as *eens pot/een spot*. By constraining the segmental content of these ambiguous phrases, we were able to carry out a detailed acoustic analysis of our stimuli and directly test whether particular aspects of acoustic-phonetic detail influence listener performance. Although the acoustic analysis revealed several differences in the realization of the two possible readings of the ambiguous phrases, only one of these differences correlated with the participants' performance in the eye-tracking task. These findings lead to two conclusions. First, finding a difference in the acoustic properties of speech stimuli is not sufficient to conclude that participants use that particular difference in lexical disambiguation. Such a conclusion needs to be based on a direct test indicating that that acoustic property modulates spoken-word recognition. Second, our findings show that individual segment duration, such as the duration of the [s] in the ambiguous phrase *one spade/once paid*, can bias listeners' interpretation of such utterances.

REFERENCES

Allen, J. S., & Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics*, **63**, 798-810.

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 & Language*, **38**, 419-439.

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and visual world. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 347-386). New York: Psychology Press.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, **52**, 163-187.

Art Explosion Library [software] (1995). Calabasas, CA: Nova Development.

Barry, W. J. (1981). Internal juncture and speech communication. In W. J. Barry & K. J. Kohler (Eds.), *Beiträge zur experimentellen und angewandten Phonetik* (pp. 229-289). Kiel, Germany: AIPUK.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. In C. Ewen & J. Anderson (Eds.), *Phonology yearbook* (Vol. 3, pp. 255-309). Cambridge: Cambridge University Press.

Cho, T. H., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, **29**, 155-190.

Cho, T. H., McQueen, J. M., & Cox, E. A. (in press). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*.

Christie, W. M. (1974). Some cues for syllable juncture perception in English. *Journal of the Acoustical Society of America*, **55**, 819-821.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory & Language*, **51**, 523-547.

Cole, R. A., & Cooper, W. E. (1975). Perception of voicing in English affricates and fricatives. *Journal of the Acoustical Society of America*, **58**, 1280-1287.

Cycowicz, Y. M., Friedman, D., Rothstein, M., & Snodgrass, J. G. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of Experimental Child Psychology*, **65**, 171-237.

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 & Cognitive Processes*, **16**, 507-534.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 218-244.

Fischer, B. (1992). Saccadic reaction time: Implications for reading, dyslexia and visual cognition. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 31-45). New York: Springer.

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, **29**, 109-135.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, **101**, 3728-3740.

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

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, **105**, 251-279.

Gow, D. W. (2002). Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 163-179.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 344-359.

Hallett, P. E. (1986). Eye movements. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 10.1-10.112). New York: Wiley.

Johnson, K. (1997a). The auditory/perceptual basis for speech segmentation. *Ohio State University Working Papers in Linguistics*, **50**, 101-113.

Johnson, K. (1997b). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145-165). San Diego: Academic Press.

Jongman, A. (1989). Duration of frication noise required for identification of English fricatives. *Journal of the Acoustical Society of America*, **85**, 1718-1725.

Kemps, R. J. J. K. (2004). *Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction*. Doctoral dissertation, Radboud University, Nijmegen (MPI Series in Psycholinguistics, Vol. 28). Wageningen: Ponsen & Looijen.

Klatt, D. (1974). Duration of [s] in English words. *Journal of Speech & Hearing Research*, **17**, 51-63.

Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. *Phonetica*, **5**, 1-54.

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

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, **53**, 372-380.

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

McQueen, J. M., Dahan, D., & Cutler, A. (2003). Continuity and gradedness in speech processing. In A. S. Meyer & N. O. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 39-78). Berlin: Mouton de Gruyter.

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.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Erlbaum.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, **25**, 457-465.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, **46**, 505-512.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, **62**, 714-719.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris.

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

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, **34**, 191-243.

Oller, D. K. (1973). Effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, **54**, 1235-1247.

Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics*, **20**, 331-350.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, **90**, 51-89.

Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America*, **57**, 1030-1033.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, **25**, 193-247.

Shatzman, K. B. (2004). Segmenting ambiguous phrases using phoneme duration. In S. H. Kin & M. J. Bae (Eds.), *Proceedings of the Eighth International Conference on Spoken Language Processing* (pp. 329-332). Seoul: Sunjin.

Snodgrass, J. G., & Vanderwart, M. (1980). Standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 174-215.

Spinelli, E., McQueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory & Language*, **48**, 233-254.

Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., & Kurowski, K. (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America*, **91**, 2979-3000.

Streeter, L. A., & Nigro, G. N. (1979). Role of medial consonant transitions in word perception. *Journal of the Acoustical Society of America*, **65**, 1533-1541.

Tabossi, P., Collina, S., Mazzetti, M., & Zoppello, M. (2000). Syllables in the processing of spoken Italian. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 758-775.

Tanenhaus, M. K., & Spivey-Knowlton, M. J. (1996). Eye-tracking. *Language & Cognitive Processes*, **11**, 583-588.

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

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, **28**, 397-440.

Umeda, N. (1977). Consonant duration in American English. *Journal of the Acoustical Society of America*, **61**, 846-858.

Waals, J. (1999). *An experimental view of the Dutch syllable.* Doctoral dissertation, Utrecht University, Utrecht (LOT Dissertation Series, Vol. 18). Utrecht: LOT.

NOTES

1. The pictures are available on request from the first author.

2. There is some variability in the eye-tracking literature regarding the lag between significant events in the speech stream and changes in fixation proportions. While many studies report a lag of about 200 msec, it is not uncommon to find values closer to 300 msec. A reviewer suggested that analysis based on fixation times (as in the present study) might overestimate the time locking of eye movements and speech. Fixation times are based on dwell times; that is, the onset of a fixation is the point in time when the eye has become relatively stationary. According to this suggestion, the initiation of a saccade might be a more appropriate

index for estimating the time locking, because the launch of a saccade nearly always indicates that the target has been selected. While there is merit in this suggestion, it does not explain the existing variability in time locking that is found in the literature, because most published studies use dwell times. To allow comparability with previous studies, we report analyses based on dwell times. We did, however, reanalyze the data of Experiment 1 by defining the onset of each fixation as the time in which the saccade preceding that fixation was initiated. In the new analysis, fixation proportions to the competitor started rising at around 250 msec after the splicing point (i.e., the average duration of saccades preceding fixations was 50 msec). This estimate is still longer than the 200 msec reported in some studies. Note, however, that Altmann and Kamide (2004) have argued that the estimation of 200 msec for saccade planning and launching is accurate when the target is known, but not when participants have to recognize the target word in the incoming speech stream in order to identify the target picture, as is the case in the eye-tracking paradigm. It seems plausible that differences among studies in the time locking of fixations and speech may therefore vary with the difficulty of recognizing the target word.

[3] Because the onsets of the target and the competitor are not aligned, a two-way (picture [target vs. competitor] × splicing condition) ANOVA would be inappropriate.

APPENDIX
Stimulus Sets Used in Experiments 1 and 2

Target	Competitor	Distractor	Distractor
pan (pan)	sprinkhaan (grasshopper)	ladder (ladder)	jurk (dress)
peen (carrot)	spier (muscle)	tafel (table)	wolk (cloud)
peer (pear)	spuit (syringe)	vlieger (kite)	boot (boat)
pier (worm)	spaan (oar)	riem (belt)	klomp (clog)
pijl (arrow)	speen (pacifier)	glas (glass)	wiel (wheel)
pil (pill)	spatel (spatula)	raket (rocket)	tomaat (tomato)
pin (pin)	speer (spear)	raam (window)	jas (jacket)
pion (pawn)	spijker (nail)	dak (roof)	banaan (banana)
pit (pit)	spook (ghost)	fototoestel (camera)	bus (bus)
pot (jar)	spin (spider)	vuur (fire)	kompas (compass)
prei (leek)	spiegel (mirror)	tent (tent)	fiets (bicycle)
taart (cake)	stuur (handlebars)	pet (cap)	boek (book)
tand (tooth)	stier (bull)	put (well)	bezem (broom)
tang (pliers)	stempel (stamp)	koffer (suitcase)	bank (sofa)
teen (toe)	strik (bow)	bal (ball)	waaier (fan)
teil (tub)	step (scooter)	panty (panty hose)	hand (hand)
tempel (temple)	strijkijzer (iron)	band (tire)	piano (piano)
thee (tea)	ster (star)	bril (glasses)	muur (wall)
tol (top)	staart (tail)	paraplu (umbrella)	boor (drill)
tulp (tulip)	stekker (plug)	laars (boot)	knoop (button)

(Manuscript received August 9, 2004;
revision accepted for publication February 4, 2005.)